HYPOTHESIS TESTING AND ANOVA

WITH PYTHON
BY DR. ALVIN ANG

CONTENTS

I. HYPOTHESIS TESTING FOR SLEEP.CSV

- The file can be found here: https://www.alvinang.sg/s/sleep.csv
- https://www.alvinang.sg/s/Hypothesis_Testing_and_ANOVA_with_Python_by_Dr_Alvin_Ang.ipynb

	A	B	C	D
1		extra	group	ID
2	1	0.7	1	1
3	2	-1.6	1	2
4	3	-0.2	1	3
5	4	-1.2	1	4
6	5	-0.1	1	5
7	6	3.4	1	6
8	7	3.7	1	7
9	8	0.8	1	8
10	9	0	1	9
11	10	2	1	10
12	11	1.9	2	1
13	12	0.8	2	2
14	13	1.1	2	3
15	14	0.1	2	4
16	15	-0.1	2	5
17	16	4.4	2	6
18	17	5.5	2	7
19	18	1.6	2	8
20	19	4.6	2	9
21	20	3.4	2	10

A. STEP 1: IMPORT THE SLEEP DATASET

Hypothesis Testing for Sleep.csv

- Do the 2 groups of people have the same amount of sleep?

Step 1: Import the Sleep Dataset

```
[ ] import statsmodels.api as sm
    sleep = sm.datasets.get_rdataset("sleep").data
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:1
  import pandas.util.testing as tm
```

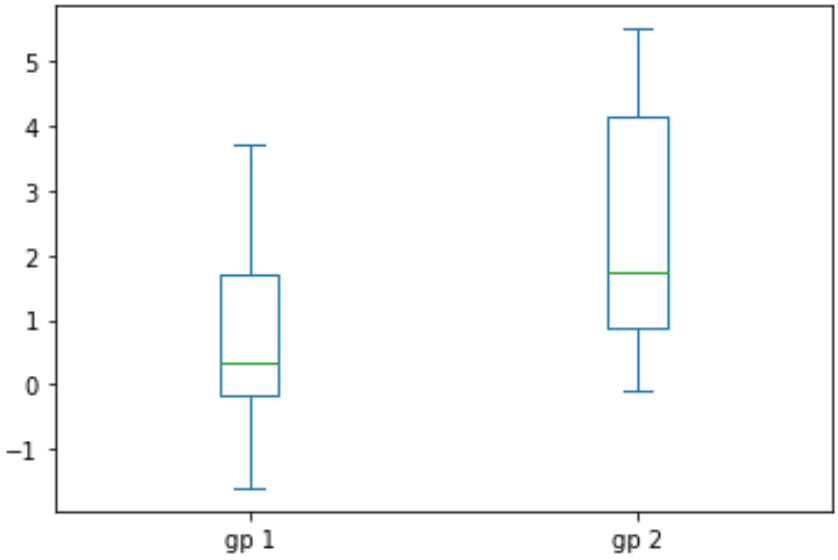
```
#preview the dataset
sleep
```

	extra	group	ID
0	0.7	1	1
1	-1.6	1	2
2	-0.2	1	3
3	-1.2	1	4
4	-0.1	1	5
5	3.4	1	6
6	3.7	1	7
7	0.8	1	8
8	0.0	1	9
9	2.0	1	10
10	1.9	2	1
11	0.8	2	2
12	1.1	2	3
13	0.1	2	4
14	-0.1	2	5
15	4.4	2	6
16	5.5	2	7
17	1.6	2	8
18	4.6	2	9
19	3.4	2	10

B. STEP 2: BOX PLOT

Step 2: Box Plot for the 2 Groups

```
import pandas as pd

gp1 = sleep[sleep.group==1].extra
gp2 = sleep[sleep.group==2].extra

c = pd.DataFrame({'gp 1': gp1.values, 'gp 2': gp2.values})

c.plot(kind = 'box')

#at first glance, both box plots look the same
#i.e. no significant difference between both groups
#because they overlap each other but we can't be sure
#so we need to do hypothesis testing to confirm this
```



C. STEP 3: INDIVIDUAL T TEST BETWEEEN BOTH GROUPS

Step 3: Individual t test Between Both Groups

```
import scipy

result = scipy.stats.ttest_ind(gp1, gp2)
result.pvalue

#the p-value = 0.079 > alpha (0.05)
#Thus we ACCEPT H0 --> Both Groups have NO Significant Difference
#they are the SAME
```

```
0.07918671421593818
```

D. STEP 4: ALTERNATIVE WAY OF RUNNING T TEST

Step 4: Alternative Way of Running t Test

```
import numpy as np
from scipy.stats import ttest_ind

def t_test(x, y, alternative = 'both-sided'):
            _, double_p = ttest_ind(x, y, equal_var = False)
            if alternative == 'both-sided':
              pval = double_p
            elif alternative == 'greater':
              if np.mean(x) > np.mean(y):
                pval = double_p/2.
              else:
                  pval = 1.0 - double_p/2.
            elif alternative == 'less':
              if np.mean(x) < np.mean(y):
                pval = double_p/2.
              else:
                  pval = 1.0 - double_p/2.
            return pval

print(t_test(gp1, gp2, alternative = 'both-sided'))

#For 2 sided test, we see that it gives the same p value of 0.079

#though for this case the result is the same as Step 3,
#we may make use of the above script to test for 'greater' or 'less'
#than hypothesis tests should there be a difference in the means.
```

```
0.0793941401873582
```

The file can be found here: https://www.alvinang.sg/s/chickwts.csv

	A	B	C
1		weight	feed
2	0	179	horsebean
3	1	160	horsebean
4	2	136	horsebean
5	3	227	horsebean
6	4	217	horsebean
7	5	168	horsebean
8	6	108	horsebean
9	7	124	horsebean
10	8	143	horsebean
11	9	140	horsebean
12	10	309	linseed
13	11	229	linseed
14	12	181	linseed
15	13	141	linseed
16	14	260	linseed
17	15	203	linseed
18	16	148	linseed
19	17	169	linseed
20	18	213	linseed
21	19	257	linseed
22	20	244	linseed

A. STEP 1: IMPORT CHICKWTS.CSV

Hypothesis Testing for Chickwts.csv

- Comparing the feeds Horsebean vs Casein, is there a significant difference in the weights of the chicken?
- If there is, which one has more?
- I.e. will Horsebean or Casein make the Chicken fatter?

Step 1: Import Chickwts.csv

```
[ ] import statsmodels.api as sm
    import scipy

    chickwts = sm.datasets.get_rdataset("chickwts").data
```

```
#preview the dataset
chickwts
```

	weight	feed
0	179	horsebean
1	160	horsebean
2	136	horsebean
3	227	horsebean
4	217	horsebean
...	...	...
66	359	casein
67	216	casein
68	222	casein
69	283	casein
70	332	casein

71 rows × 2 columns

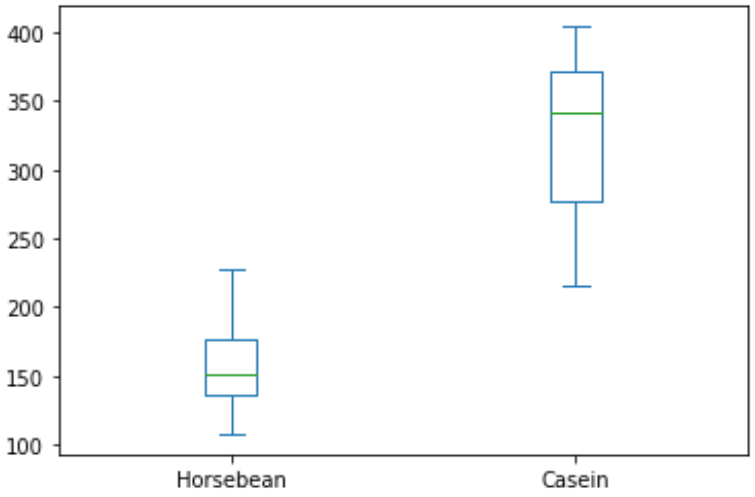
B. STEP 2: EXTRACT ONLY THE WEIGHTS COLUMN FOR HORSEBEAN AND CASEIN

Step 2: Extract Only the Weights Column for Horsebean and Casein

```
[ ] #reading in the weights for the 2 type of feed
    horsebean = chickwts[chickwts.feed == 'horsebean'].weight
    casein = chickwts[chickwts.feed=='casein'].weight
```

```
import pandas as pd

c = pd.DataFrame({'Horsebean': horsebean, 'Casein': casein})

c.plot(kind = 'box')

#at first glance, both box plots look DIFFERENT
#i.e. SIGNIFICANT difference between both groups
#because they DON'T overlap each other but we can't be sure
#so we need to do hypothesis testing to confirm this
```

```
/usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarn
  X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
<matplotlib.axes._subplots.AxesSubplot at 0x7f68f3438790>
```



C. STEP 3: PERFORM A 2 TAIL T - TEST

Step 3: Perform a 2 tail t - test

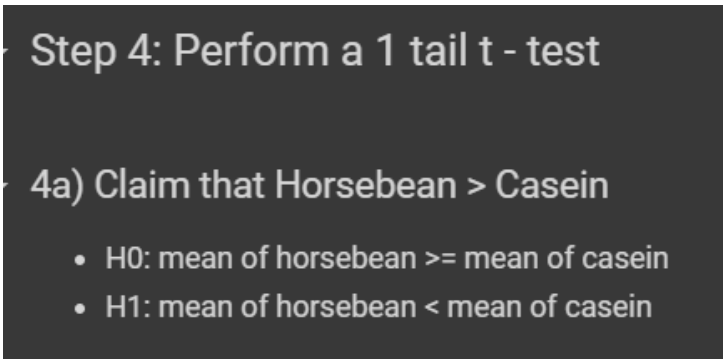
- H0: Horsebean == Casein
- H1: Horsebean != Casein

```
#t test 2 tail
result = scipy.stats.ttest_ind(horsebean, casein)
result.pvalue

#P - value = 0.0000... which means we accept H1
#Meaning, there is a SIGNIFICANT difference
#between 'horsebean' vs ' casein' feeds.
#But which is more? Which is lesser?
```

```
8.254541016953191e-07
```

D. STEP 4: PERFORM A 1 TAIL T – TEST

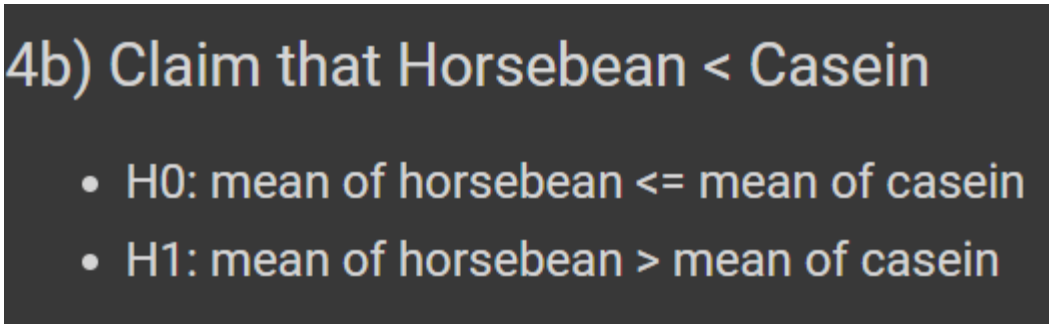
1. CLAIM THAT HORSEBEAN > CASEIN




```
import numpy as np
from scipy.stats import ttest_ind

def t_test(x, y, alternative = 'both-sided'):
            _, double_p = ttest_ind(x, y, equal_var = False)
            if alternative == 'both-sided':
              pval = double_p
            elif alternative == 'greater':
              if np.mean(x) > np.mean(y):
                pval = double_p/2.
              else:
                  pval = 1.0 - double_p/2.
            elif alternative == 'less':
              if np.mean(x) < np.mean(y):
                pval = double_p/2.
              else:
                  pval = 1.0 - double_p/2.
            return pval

print(t_test(horsebean, casein, alternative = 'less'))

#The test here is:
# - H0: mean of horsebean >= mean of casein
# - H1: mean of horsebean < mean of casein

#P value is very small << alpha (0.05) --> Accept H1
#Meaning, the mean weight of horsebean is significantly less than casein.
```

```
3.60512479059563e-07
```

2. CLAIM THAT HORSEBEAN < CASEIN




```
[ ] import numpy as np
    from scipy.stats import ttest_ind

    def t_test(x, y, alternative = 'both-sided'):
              _, double_p = ttest_ind(x, y, equal_var = False)
              if alternative == 'both-sided':
                pval = double_p
              elif alternative == 'greater':
                if np.mean(x) > np.mean(y):
                  pval = double_p/2.
                else:
                    pval = 1.0 - double_p/2.
              elif alternative == 'less':
                if np.mean(x) < np.mean(y):
                  pval = double_p/2.
                else:
                    pval = 1.0 - double_p/2.
              return pval

    print(t_test(horsebean, casein, alternative = 'greater'))

    #The test here is:
    # - H0: mean of horsebean <= mean of casein
    # - H1: mean of horsebean > mean of casein

    #P value is almost 1, very big >> alpha (0.05) --> Accept H0
    #Meaning, the mean weight of horsebean is significantly less than casein.

    0.9999996394875209
```

The file can be found here: https://www.alvinang.sg/s/chickwts.csv

A. STEP 1: IMPORT CHICKWTS DATASET

ANOVA for Chickwts.csv

Step 1: Import Chickwts Dataset

```
[2] import statsmodels.api as sm
    chickwts = sm.datasets.get_rdataset("chickwts").data
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: Fu
  import pandas.util.testing as tm
```

```
#preview the data
chickwts
```

	weight	feed
0	179	horsebean
1	160	horsebean
2	136	horsebean
3	227	horsebean
4	217	horsebean
...	...	...
66	359	casein
67	216	casein
68	222	casein
69	283	casein
70	332	casein

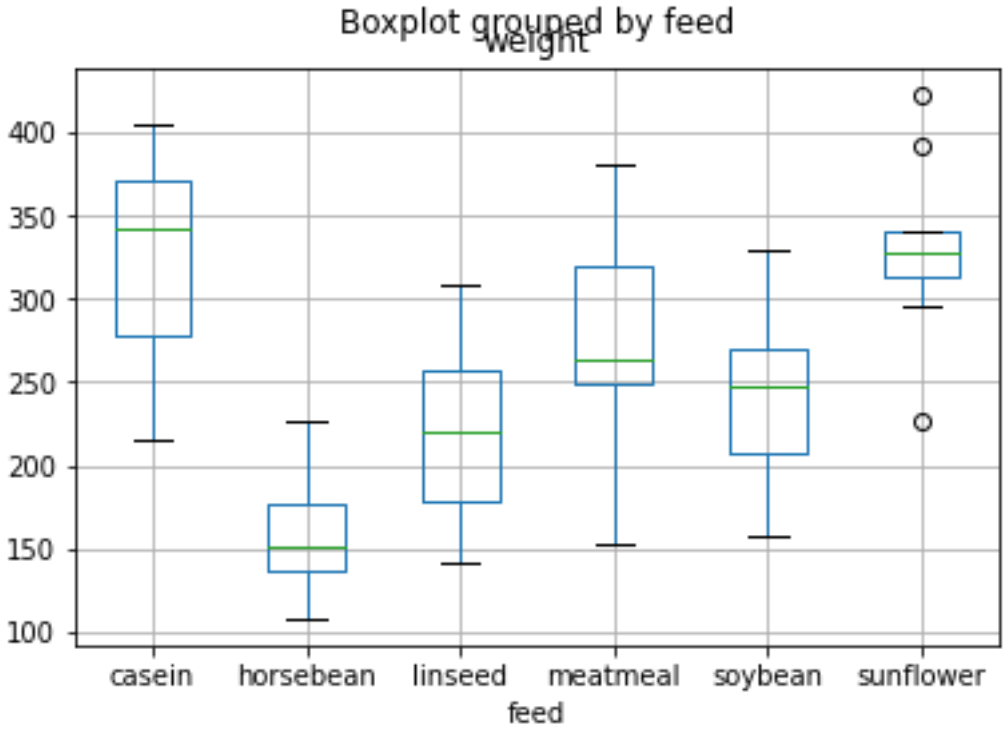
71 rows × 2 columns

B. STEP 2: BOXPLOT

Step 2: Boxplot

```
chickwts.boxplot('weight', 'feed')

#observations:
#most significant difference is between Horsebean vs Casein
#thus the initial claim will be
#--> There IS A SIGNIFICANT difference in at least two of the feeds
```

```
/usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarni
  X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
<matplotlib.axes._subplots.AxesSubplot at 0x7f067b68ab50>
```



C. STEP 3: ANOVA TEST

Step 3: ANOVA Test

```
[6] from statsmodels.formula.api import ols

    model = ols('weight ~ feed', chickwts).fit()
```

```
from statsmodels.stats.api import anova_lm

anova_lm(model)

#H0: Mean Weight for Casein = Horsebean = .....= Sunflower
#H1: Mean Weight for Casein != Horsebean != ..... != Sunflower

#P value = 0.0000... << Alpha (0.05) --> ACCEPT H1
#--> There IS A SIGNIFICANT difference in at least two of the feeds
```

	df	sum_sq	mean_sq	F	PR(>F)
feed	5.0	231129.162103	46225.832421	15.3648	5.936420e-10
Residual	65.0	195556.020996	3008.554169	NaN	NaN

IV. ANOVA FOR COLLEGE.CSV – PART I: IS THERE A SIGNIFICANT DIFFERENCE IN TUITION NUMBERS BETWEEN REGIONS?

The file can be found here: https://www.alvinang.sg/s/college.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	name	city	state	region	highest_degree	control	gender	admission_rate	sat_avg	undergrads	tuition	faculty_salary_avg	loan_default_rate	median_debt	lon	lat
2	102669	Alaska Pacific University	Anchorage	AK	West	Graduate	Private	CoEd	0.4207	1054	275	19610	5804	0.077	23250	-149.9002778	61.2180556
3	101648	Marion Military Institute	Marion	AL	South	Associate	Public	CoEd	0.6139	1055	433	8778	5916	0.136	11500	-87.3191655	32.6323536
4	100830	Auburn University at Montgomery	Montgomery	AL	South	Graduate	Public	CoEd	0.8017	1009	4304	9080	7255	0.106	21335	-86.2999689	32.3668052
5	101879	University of North Alabama	Florence	AL	South	Graduate	Public	CoEd	0.6788	1029	5485	7412	7424	0.111	21500	-87.677251	34.79981
6	100858	Auburn University	Auburn	AL	South	Graduate	Public	CoEd	0.8347	1215	20514	10200	9487	0.045	21831	-85.4807825	32.6098566
7	100663	University of Alabama at Birmingham	Birmingham	AL	South	Graduate	Public	CoEd	0.8569	1107	11383	7510	9957	0.062	21941.5	-86.80249	33.5206608
8	101480	Jacksonville State University	Jacksonville	AL	South	Graduate	Public	CoEd	0.8326	1041	7060	7092	6801	0.096	23000	-85.7613536	33.8137125
9	102049	Samford University	Birmingham	AL	South	Graduate	Private	CoEd	0.5954	1165	3033	27324	8367	0.007	23000	-86.80249	33.5206608
10	101709	University of Montevallo	Montevallo	AL	South	Graduate	Public	CoEd	0.743	1070	2644	10660	7437	0.103	23266	-86.8641558	33.1006746
11	100751	The University of Alabama	Tuscaloosa	AL	South	Graduate	Public	CoEd	0.5105	1185	29851	9826	9667	0.063	23750	-87.5691735	33.2098407
12	102261	Southeastern Bible College	Birmingham	AL	South	Bachelor	Private	CoEd	1	930	170	11370	4554	0.048	24000	-86.80249	33.5206608
13	100706	University of Alabama in Huntsville	Huntsville	AL	South	Graduate	Public	CoEd	0.8203	1219	5451	9158	9302	0.061	24097	-86.5861037	34.7303688
14	101587	University of West Alabama	Livingston	AL	South	Graduate	Public	CoEd	0.7199	990	1916	8018	6146	0.078	24253	-88.1872475	32.5843025
15	102094	University of South Alabama	Mobile	AL	South	Graduate	Public	CoEd	0.8335	1048	11267	7188	7195	0.075	24711	-88.0398912	30.6953657
16	102368	Troy University	Troy	AL	South	Graduate	Public	CoEd	0.4414	1050	15025	7564	6246	0.114	25000	-85.969951	31.8087678
17	101435	Huntingdon College	Montgomery	AL	South	Bachelor	Private	CoEd	0.5839	1026	1149	24550	5772	0.102	26230	-86.2999689	32.3668052
18	101693	University of Mobile	Mobile	AL	South	Graduate	Private	CoEd	0.5847	1014	1460	19475	4914	0.062	27000	-88.0398912	30.6953657
19	102234	Spring Hill College	Mobile	AL	South	Graduate	Private	CoEd	0.5177	1116	1215	32468	6071	0.066	27000	-88.0398912	30.6953657
20	100937	Birmingham Southern College	Birmingham	AL	South	Bachelor	Private	CoEd	0.5339	1181	1180	31708	7451	0.044	27000	-86.80249	33.5206608
21	101912	Oakwood University	Huntsville	AL	South	Graduate	Private	CoEd	0.4787	928	1878	16720	5147	0.125	27250	-86.5861037	34.7303688
22	101073	Concordia College Alabama	Selma	AL	South	Bachelor	Private	CoEd	0.5328	942	322	10320	5812	0.315	32000	-87.0211007	32.4073589
23	100724	Alabama State University	Montgomery	AL	South	Graduate	Public	CoEd	0.5326	851	4811	8720	6609	0.156	33118.5	-86.2999689	32.3668052
24	102377	Tuskegee University	Tuskegee	AL	South	Graduate	Private	CoEd	0.4922	978	2588	19570	8399	0.128	33500	-85.7077266	32.430237
25	100654	Alabama A & M University	Normal	AL	South	Graduate	Public	CoEd	0.5256	827	4206	9096	6892	0.172	33888	-86.5722237	34.7838409
26	102270	Stillman College	Tuscaloosa	AL	South	Bachelor	Private	CoEd	0.5901	811	1056	15865	4597	0.187	38218	-87.5691735	33.2098407
27	107585	University of Arkansas Community College-Morrilton	Morrilton	AR	South	Associate	Public	CoEd	0.6181	930	1920	2732	4855	0.169	8000.5	-92.7440538	35.1509173
28	107983	Southern Arkansas University Main Campus	Magnolia	AR	South	Graduate	Public	CoEd	0.7142	989	2784	7736	6269	0.18	17000	-93.239334	33.2670725
29	106467	Arkansas Tech University	Russellville	AR	South	Graduate	Public	CoEd	0.8626	1010	8845	5862	6083	0.171	17480	-93.1337856	35.2784173
30	107877	Williams Baptist College	Walnut Ridge	AR	South	Bachelor	Private	CoEd	0.7049	983	508	14360	4720	0.089	19000	-90.9559534	36.0684035
31	106458	Arkansas State University-Main Campus	Jonesboro	AR	South	Graduate	Public	CoEd	0.7239	1088	9139	7720	6927	0.102	19250	-90.704279	35.8422967
32	107071	Henderson State University	Arkadelphia	AR	South	Graduate	Public	CoEd	0.6278	989	3226	7860	5624	0.149	19586	-93.0537839	34.1209292
33	106713	Central Baptist College	Conway	AR	South	Bachelor	Private	CoEd	0.5697	985	788	13800	4847	0.109	21500	-92.4421011	35.0886963
34	106704	University of Central Arkansas	Conway	AR	South	Graduate	Public	CoEd	0.9426	1049	9232	7889	6562	0.091	21500	-92.4421011	35.0886963

A. STEP 1: IMPORTING COLLEGE.CSV

ANOVA for College.csv - Part I: Is there a Significant Difference in Tuition Numbers between Regions?

Step 1: Importing College.csv

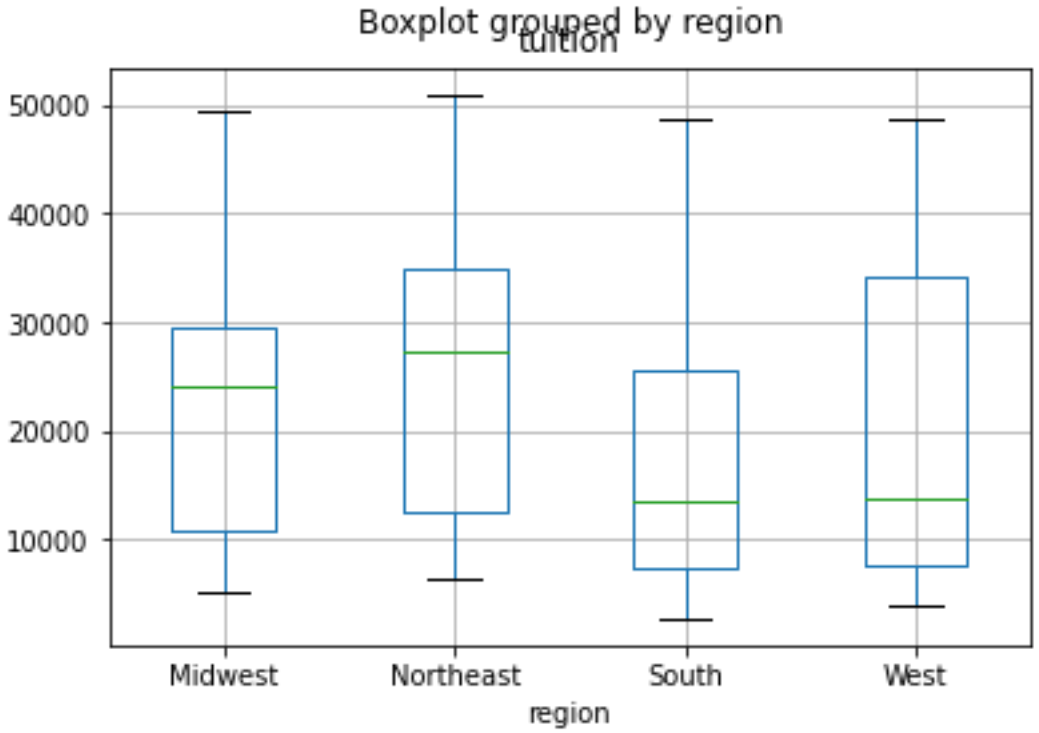
```
[8] import pandas as pd

    college = pd.read_csv('https://www.alvinang.sg/s/college.csv')
```

B. STEP 2: BOX PLOT

Step 2: Box Plot

```
college.boxplot('tuition', 'region')

#there don't seem to be a significant difference between 4 regions...
#Thus we claim NO difference...
```

```
/usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarning:
  X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
<matplotlib.axes._subplots.AxesSubplot at 0x7f068bcd3350>
```



C. STEP 3: ANOVA TEST

Step 3: ANOVA Test

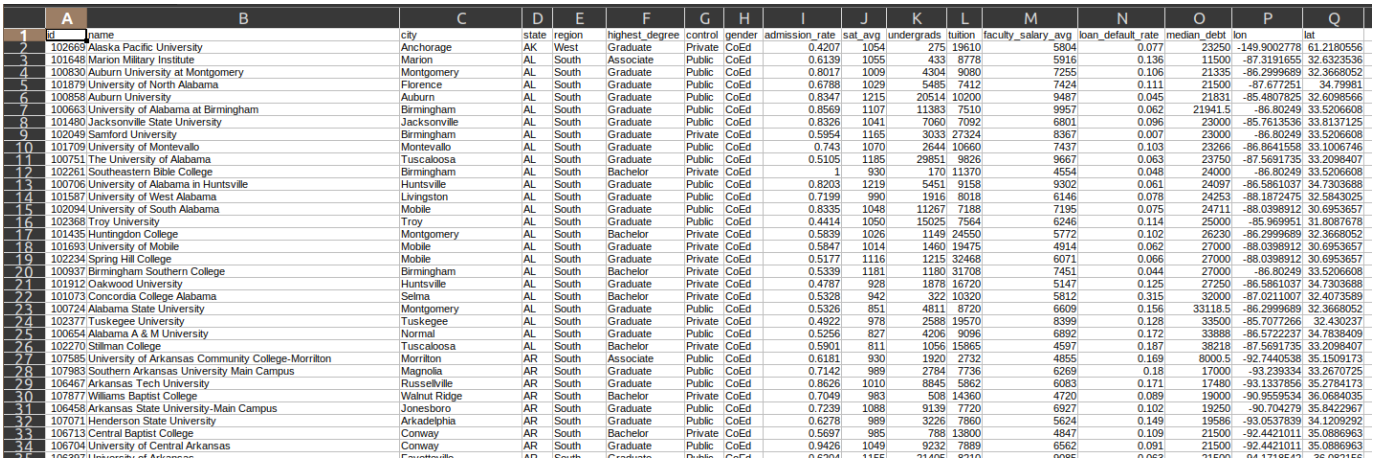
```
[10] from statsmodels.formula.api import ols

     model = ols('tuition ~ region', college).fit()
```

```
from statsmodels.stats.api import anova_lm

anova_lm(model)

#H0: Tuition Numbers in Midwest = Northeast = South = West
#H1: Tuition Numbers in Midwest != Northeast != South != West

#P value = 0.000.... << Alpha (0.05) --> Accept H1
#There's a SIGNIFICANT difference

#Our initial observation (from box plot) is WRONG!
```

	df	sum_sq	mean_sq	F	PR(>F)
region	3.0	1.240011e+10	4.133370e+09	27.933294	1.719599e-17
Residual	1265.0	1.871857e+11	1.479729e+08	NaN	NaN

V. ANOVA FOR COLLEGE.CSV – PART II: IS THERE A SIGNIFICANT DIFFERENCE IN TUITION NUMBERS BETWEEN PRIVATE / PUBLIC SCHOOLS?

The file can be found here: https://www.alvinang.sg/s/college.csv



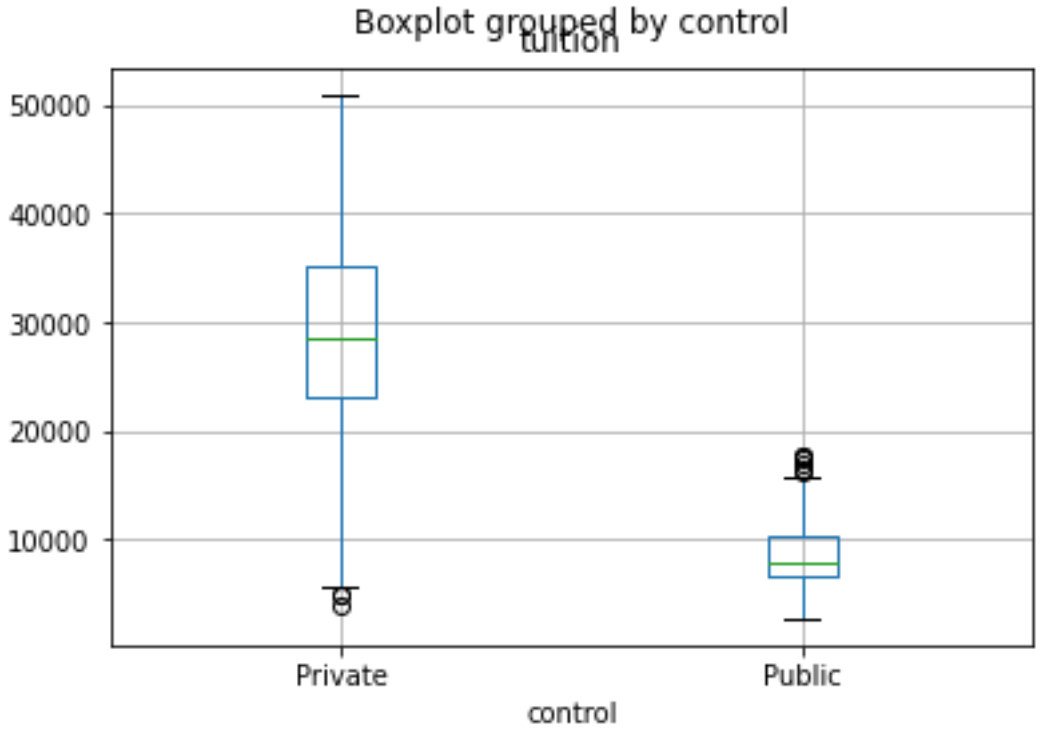
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	name	city	state	region	highest_degree	control	gender	admission_rate	sat_avg	undergrads	tuition	faculty_salary_avg	loan_default_rate	median_debt	lon	lat
2	102669	Alaska Pacific University	Anchorage	AK	West	Graduate	Private	CoEd	0.4207	1054	275	19610	5804	0.077	23250	-149.9002778	61.2180556
3	101648	Marion Military Institute	Marion	AL	South	Associate	Public	CoEd	0.6139	1055	433	8778	5916	0.136	11500	-87.3191655	32.6323536
4	100830	Auburn University at Montgomery	Montgomery	AL	South	Graduate	Public	CoEd	0.8017	1009	4304	9080	7255	0.106	21335	-86.2999689	32.3668052
5	101879	University of North Alabama	Florence	AL	South	Graduate	Public	CoEd	0.6788	1029	5485	7412	7424	0.111	21500	-87.677251	34.79981
6	100858	Auburn University	Auburn	AL	South	Graduate	Public	CoEd	0.8347	1215	20514	10200	9487	0.045	21831	-85.4807825	32.6098566
7	100663	University of Alabama at Birmingham	Birmingham	AL	South	Graduate	Public	CoEd	0.8569	1107	11383	7510	9957	0.062	21941.5	-86.80249	33.5206608
8	101480	Jacksonville State University	Jacksonville	AL	South	Graduate	Public	CoEd	0.8326	1041	7060	7092	6801	0.096	23000	-85.7613536	33.8137125
9	102049	Samford University	Birmingham	AL	South	Graduate	Private	CoEd	0.5954	1165	3033	27324	8367	0.007	23000	-86.80249	33.5206608
10	101709	University of Montevallo	Montevallo	AL	South	Graduate	Public	CoEd	0.743	1070	2644	10660	7437	0.103	23266	-86.8641558	33.1006746
11	100751	The University of Alabama	Tuscaloosa	AL	South	Graduate	Public	CoEd	0.5105	1185	29851	9826	9667	0.063	23750	-87.5691735	33.2098407
12	102261	Southeastern Bible College	Birmingham	AL	South	Bachelor	Private	CoEd	1	930	170	11370	4554	0.048	24000	-86.80249	33.5206608
13	100706	University of Alabama in Huntsville	Huntsville	AL	South	Graduate	Public	CoEd	0.8203	1219	5451	9158	9302	0.061	24097	-86.5861037	34.7303688
14	101587	University of West Alabama	Livingston	AL	South	Graduate	Public	CoEd	0.7199	990	1916	8018	6146	0.078	24253	-88.1872475	32.5843025
15	102094	University of South Alabama	Mobile	AL	South	Graduate	Public	CoEd	0.8335	1048	11267	7188	7195	0.075	24711	-88.0398912	30.6953657
16	102368	Troy University	Troy	AL	South	Graduate	Public	CoEd	0.4414	1050	15025	7564	6246	0.114	25000	-85.969951	31.8087678
17	101435	Huntingdon College	Montgomery	AL	South	Bachelor	Private	CoEd	0.5839	1026	1149	24550	5772	0.102	26230	-86.2999689	32.3668052
18	101693	University of Mobile	Mobile	AL	South	Graduate	Private	CoEd	0.5847	1014	1460	19475	4914	0.062	27000	-88.0398912	30.6953657
19	102234	Spring Hill College	Mobile	AL	South	Graduate	Private	CoEd	0.5177	1116	1215	32468	6071	0.066	27000	-88.0398912	30.6953657
20	100937	Birmingham Southern College	Birmingham	AL	South	Bachelor	Private	CoEd	0.5339	1181	1180	31708	7451	0.044	27000	-86.80249	33.5206608
21	101912	Oakwood University	Huntsville	AL	South	Graduate	Private	CoEd	0.4787	928	1878	16720	5147	0.125	27250	-86.5861037	34.7303688
22	101073	Concordia College Alabama	Selma	AL	South	Bachelor	Private	CoEd	0.5328	942	322	10320	5812	0.315	32000	-87.0211007	32.4073589
23	100724	Alabama State University	Montgomery	AL	South	Graduate	Public	CoEd	0.5326	851	4811	8720	6609	0.156	33118.5	-86.2999689	32.3668052
24	102377	Tuskegee University	Tuskegee	AL	South	Graduate	Private	CoEd	0.4922	978	2588	19570	8399	0.128	33500	-85.7077266	32.430237
25	100654	Alabama A & M University	Normal	AL	South	Graduate	Public	CoEd	0.5256	827	4206	9096	6892	0.172	33888	-86.5722237	34.7838409
26	102270	Stillman College	Tuscaloosa	AL	South	Bachelor	Private	CoEd	0.5901	811	1056	15865	4597	0.187	38218	-87.5691735	33.2098407
27	107585	University of Arkansas Community College-Morrilton	Morrilton	AR	South	Associate	Public	CoEd	0.6181	930	1920	2732	4855	0.169	8000.5	-92.7440538	35.1509173
28	107983	Southern Arkansas University Main Campus	Magnolia	AR	South	Graduate	Public	CoEd	0.7142	989	2784	7736	6269	0.18	17000	-93.239334	33.2670725
29	106467	Arkansas Tech University	Russellville	AR	South	Graduate	Public	CoEd	0.8626	1010	8845	5862	6083	0.171	17480	-93.1337856	35.2784173
30	107877	Williams Baptist College	Walnut Ridge	AR	South	Bachelor	Private	CoEd	0.7049	983	508	14360	4720	0.089	19000	-90.9559534	36.0684035
31	106458	Arkansas State University-Main Campus	Jonesboro	AR	South	Graduate	Public	CoEd	0.7239	1088	9139	7720	6927	0.102	19250	-90.704279	35.8422967
32	107071	Henderson State University	Arkadelphia	AR	South	Graduate	Public	CoEd	0.6278	989	3226	7860	5624	0.149	19586	-93.0537839	34.1209292
33	106713	Central Baptist College	Conway	AR	South	Bachelor	Private	CoEd	0.5697	985	788	13800	4847	0.109	21500	-92.4421011	35.0886963
34	106704	University of Central Arkansas	Conway	AR	South	Graduate	Public	CoEd	0.9426	1049	9232	7889	6562	0.091	21500	-92.4421011	35.0886963

A. STEP 1: BOX PLOT

Step 1: Box Plot

```
college.boxplot('tuition', 'control')

#box plot looks like there's SIGNIFICANT difference
```

```
/usr/local/lib/python3.7/dist-packages/matplotlib/cbook/__init__.py:1376: V
  X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
<matplotlib.axes._subplots.AxesSubplot at 0x7f067aedda90>
```



B. STEP 2: ANOVA TEST

Step 2: ANOVA Test

```
[13] model = ols('tuition ~ control', college).fit()
     anova_lm(model)

     #H0: Tuition Numbers for Private School = Public School
     #H1: Tuition Numbers for Private School != Public School

     #P value = 0.000..... <<Alpha (0.05) --> Accept H1
     #there is a SIGNIFICANT difference between Private vs Public school
```

	df	sum_sq	mean_sq	F	PR(>F)
control	1.0	1.262551e+11	1.262551e+11	2181.423086	9.464090e-278
Residual	1267.0	7.333068e+10	5.787741e+07	NaN	NaN

THE END

ABOUT DR. ALVIN ANG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.