Aron Chen

10/9/2022

Professor Bradford Dykes

STA 216 – Intermediate Stats

Growth As a Data Person

As the semester comes to an end I'm proud to be able to show my growth as a data person throughout this entire semester in this essay. It was a real struggle to finish this class but I'm glad to be able to finish and turn in who I am as a data person.

When it comes to demonstrating statistical programming skills I decided to go a different route instead of using SAS for the class project. I struggled with SAS for the data analysis project, so I decided to branch off and grow on my own using data frames in Python. However, I was able to create a numerical table of summary just to show I can use SAS, but I didn't include it in my project.

Substance Abuse

The MEANS Procedure

Analysis Variable : substanceabuse							
N	Minimum	Maximum	Mean	Std Dev	Median	Skewness	Kurtosis
242	0.00	1.00	0.31	0.47	0.00	0.81	-1.36

AGE

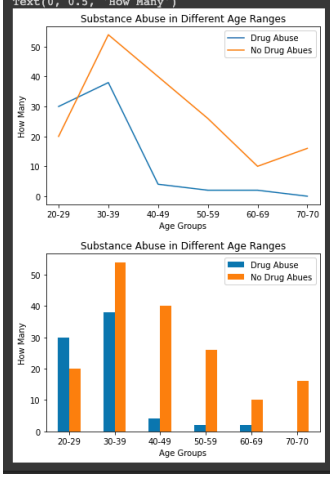
The MEANS Procedure

Analysis Variable : AGE							
N	Minimum	Maximum	Mean	Std Dev	Median	Skewness	Kurtosis
242	20.00	79.00	40.88	14.79	37.00	0.91	0.20

```
from pandas.plotting import boxplot
plot = pd.DataFrame({
    'Drug Abuse':[ages20yes,ages30yes,ages40yes,ages50yes,ages60yes,ages70yes],
    'No Drug Abues':[ages20no,ages30no,ages40no,ages50no,ages60no,ages70no]}, index=["20-29","30-39","40-49","50-59","60-69","70-70"])
lines = plot.plot.line()
plot2 = pd.DataFrame({
    'Drug Abuse':[ages20yes,ages30yes,ages40yes,ages50yes,ages60yes,ages70yes],
    'No Drug Abues':[ages20no,ages30no,ages40no,ages50no,ages60no,ages70no]},index=["20-29","30-39","40-49","50-59","60-69","70-70"])
plot3 = pd.DataFrame({
    'Drug Abuse':[ages20yes,ages30yes,ages40yes,ages50yes,ages60yes,ages70yes],
    'No Drug Abues':[ages20no,ages30no,ages40no,ages50no,ages60no,ages70no]})
ab = plot2.plot.bar(rot = 0)
ab.set_title("Substance Abuse in Different Age Ranges")
ab.set_xlabel("Age Groups")
ab.set_ylabel("How Many")
lines.set_title("Substance Abuse in Different Age Ranges")
lines.set_xlabel("Age Groups")
lines.set_ylabel("How Many")
```

Loading...

```
Text(0, 0.5, 'How Many')
```



This artifact of my source code shows that the output given is the same output we would see when using SAS properly for a line/bar graph.

Early in the semester, I had a hard time tying two variables together and be able to tell what they mean for my project. As time progressed, I was able to create a coefficient correlation table to help me produce and interpret numerical summaries of data.

```
df.corr()
```

	CLIENT_KEY	AGE	INCOME	NIGHTS	substanceabuse	completed	probation	required
CLIENT_KEY	1.000000	0.019951	-0.159723	-0.004372	-0.108912	-0.020247	0.100226	-0.037018
AGE	0.019951	1.000000	0.016940	-0.638810	-0.400848	0.332826	-0.403322	-0.262026
INCOME	-0.159723	0.016940	1.000000	-0.068616	0.112165	0.088181	-0.022632	-0.104517
NIGHTS	-0.004372	-0.638810	-0.068616	1.000000	0.458512	-0.337638	0.409137	0.273281
substanceabuse	-0.108912	-0.400848	0.112165	0.458512	1.000000	-0.235029	0.210841	0.227647
completed	-0.020247	0.332826	0.088181	-0.337638	-0.235029	1.000000	-0.325032	-0.134302
probation	0.100226	-0.403322	-0.022632	0.409137	0.210841	-0.325032	1.000000	0.125883
required	-0.037018	-0.262026	-0.104517	0.273281	0.227647	-0.134302	0.125883	1.000000

In this table between age and substance abuse we have an output of -0.4 which means there is little to no correlation between age and substance abuse. As well as a negative correlation between the two variables.

Explaining the role of variability in a variety of settings was not my strongest suit during the midterm for my project. As I did more research for my project I was able to depict what each definition means in my scenario. For example, AGE kurtosis is greater than 0 at 0.2 which means that it has a peaked shape with heavy tails, which is called leptokurtic. There are more extreme values than in a normal distribution, and the data is more concentrated around the mean.

	AGE	substanceabuse
count	242.000000	242.000000
mean	40.876033	0.314050
std	14.791256	0.465098
min	20.000000	0.000000
25%	30.000000	0.000000
50%	37.000000	0.000000
75%	49.000000	1.000000
max	79.000000	1.000000

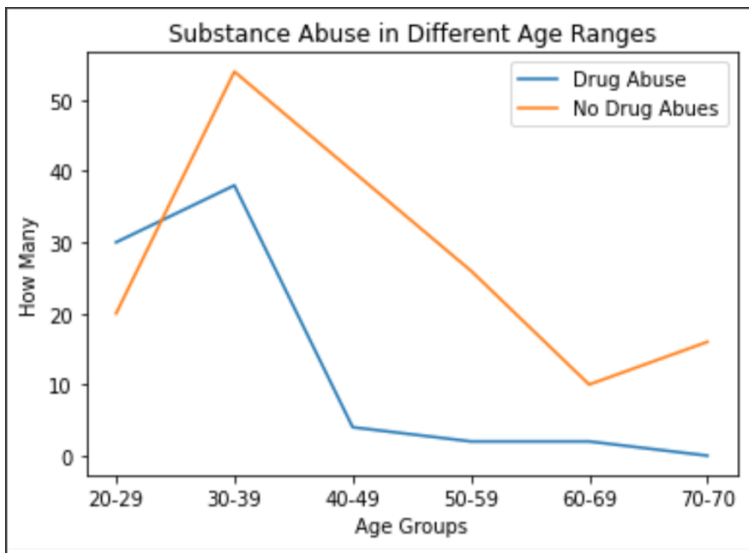
```
Age Kurtosis:  0.2006882793018363
Age Skewness:  0.9116118614992007
Substance Abuse Kurtosis:  -1.3612291325619579
Substance Abuse Skewness:  0.8062813825719001
```

When I first started the data analysis project I was struggling with creating a model to fit my data. I first started with linear regression and failed since my data was not suitable for it. I then switched over to one-way ANOVA to get the results I needed to describe my distribution of variables in a variety of settings I chose fit.

	df	sum_sq	mean_sq	F	PR(>F)
substanceabuse	1.0	8472.006420	8472.006420	45.945427	9.350906e-11
Residual	240.0	44254.274572	184.392811	NaN	NaN

During the week we started hypothesis testing I had forgotten there was even hypothesis testing in stats. So, there was only room to grow. Using my one-way ANOVA output above we can see that the p-value is significantly greater than our F-value which allows me to confidently fail to reject the null hypothesis (H0) meaning there is no difference in ages when it comes to substance abuse in homelessness. Through my hypothesis testing I have not identified a relationship between age and substance abuse.

Producing and interpreting graphical summaries of data was not very hard for me since I learned about it in STA 215. However, using a graphical summary of data to match an output of a numerical summary of data is where I had room to grow. I created this line graph to visually show that the age group 30-39 peaks with the amount of substance abuse and slowly declines as we move further up the age groups just as my correlation coefficient model predicted.



When we first started doing confidence intervals I was not confident on the concept especially when I was using python to try and create one for my project. As I worked more on the activities I was able to create a confidence interval for my AGE and substanceabuse variables. A confidence variable of (42,47) for no substanceabuse, (30,34) with substanceabuse.

With both means falling in-between our intervals we can use this sample data to make inferences about the population.

```
                      mean  count        std
substanceabuse
0                44.879518    166  15.183242
1                32.131579     76   9.104346
------------------------------
                      mean  count        std    ci95_hi    ci95_lo
substanceabuse
0                44.879518    166  15.183242  47.189276  42.569760
1                32.131579     76   9.104346  34.178486  30.084672
```