Stats 216

Aron Chen

Professor Bradford

11/15/22

Final Grade Reflection

The grade I believe I earned in this class is an A-. I think I deserve this grade because I have clearly demonstrated that I have the knowledge of most course objectives in this class. Let me properly show you with these artifacts and an explanation for each course objective.

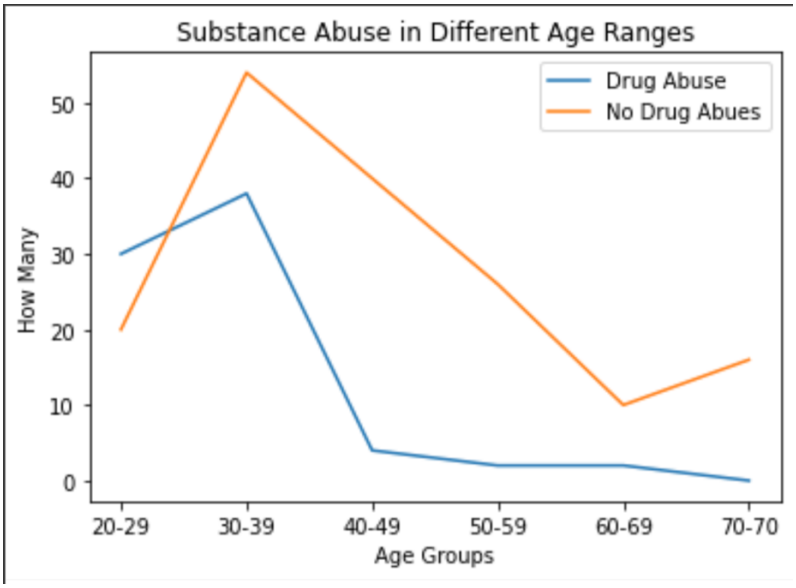
In my project I clearly demonstrated that I could create and interpret numerical summaries of data. Before creating my linear regression model and fit my data to it I created a coefficient correlation table to predict how my linear regression model would look like compared to the actual output from the model. As we can see when it comes to Age and substanceabuse there is a negative linear coefficient correlation of -0.04 which means that when as a homeless person age increases the likely hood of substance abuse decreases.

```
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

When it comes to creating graphical summaries of data, I demonstrated that quite well within my project. The first graph I created is a line graph that shows the amount of substance abuse vs nonsubstance abuse in each homeless person's age group. We can see that the age

group 30-39 peaks with the amount of substance abuse and slowly declines as we move further up in the age groups just like how my correlation coefficient model predicted.



At first, I struggled to explain the role of variability in a variety of different settings because I didn't know all the definitions. But after my project I can say that I have improved a lot. For example, we can see that our 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
```

Creating a statistical model was the hardest part for me. Originally, I was going to create a multiple linear regression model to fit my data. As the semester went on and we learned about ANOVA I realized that the best model to fit my analysis question was a one-way ANOVA test because one of my variables was a categorial which would have made my linear regression model look funky.

	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

Before this class I didn't know what hypothesis testing was in a statistical setting. After learning about one-way ANOVA and the steps to necessary to create the model I learned how to fail to reject the null hypothesis (H0). Since the p-value I got is significantly larger than our F value, I can confidently fail to reject the null hypothesis meaning there is no difference in ages when it comes to substance abuse in homelessness because my testing has not identified a relationship between age and substanceabuse.

I'm not positive that I met the criteria for the interval estimation course objective. I did struggle with this trying to figure out the code for it in python. What I did was create a high and low confidence level using substanceabuse and AGEs mean, count, and standard deviation. I got the output of Confidence interval of (42,47) no abuse and (30,34) yes to abuse. Since both our means fall in-between the intervals I think it means we can use this sample data to make inferences to 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
```

What I struggled with the most in this class was SAS. I did the activities in SAS but ultimately decided to do the project in python. Doing stats programming in python was a learning curve for me as well. I've demonstrated that I know how to import and extrapolate the correct data to support my statistical analysis project while also proving I know how to do the other learning objectives through this alley. However, I did create a numerical summary of data for substanceabuse and AGE in SAS which matches my output I created in python.

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