Final Exam

NAME:

MAT 128, SPRING 2019

Each question is worth 10 points. You are allowed one 8 1/2" x 11" sheet of paper with hand-written notes on both sides.

1. (a) The CSV file `ufo.csv` contains information about the number of reported sightings of UFOs in New York, New Jersey, Pennsylvania, and Connecticut in the last five years. Missing data is indicated with a '-1'. Write a **complete** Python program (excluding import statements) that plots the number of UFO sightings in New York over time. Your program should explicitly read in the reading data as NaN, the Pandas symbol for missing data. Your graph should have a title and axes labels.

   `ufo.csv`:

   ```
   Year,Connecticut,New Jersey,New York,Pennsylvania
   2014,4,10,-1,7
   2015,5,15,3,7
   2016,7,13,4,9
   2017,9,12,5,-1
   2018,6,10,3,6
   ```

   Answer:

   ```
   traffic = pd.read_csv("ufo.csv",na_values = "-1")
   traffic.plot(x = "Year",y="New York")
   plt.title("UFO reported sightings in New York state")
   plt.xlabel("Year")
   plt.ylabel("# of reported sightings")
   ```

   (b) For the following data, state whether the data is:

   - qualitative or quantitative
   - discrete or continuous (if applicable)
   - nominal, ordinal, interval, or ratio

   (i) Ratings (poor, average, good, great) for 100 randomly selected New York restaurants.

   Answer: qualitative,ordinal

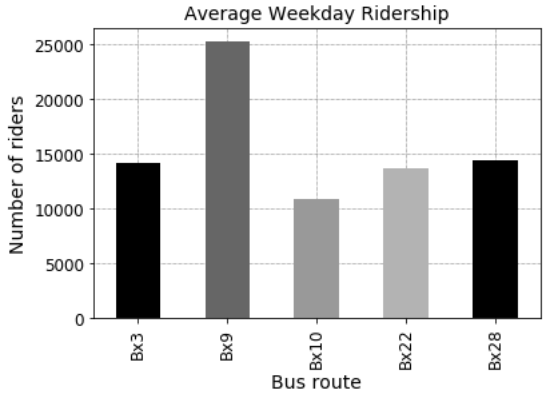
   (ii) Number of days of sunshine each year in 20 major US cities.

   Answer: quantitative, discrete, ratio

   (iii) Weights (in lbs) of dogs at a dog show.

   Answer: quantitative, continuous, ratio

2. (a) Below is a bar plot of the average weekday ridership of several bus routes in the Bronx.



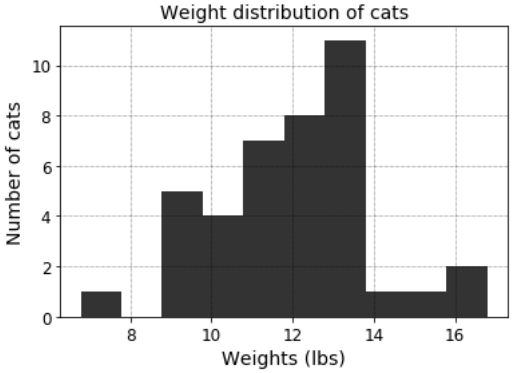
(i) Which bus route has the lowest average weekday ridership?

Answer: Bx10

(ii) How many people ride the Bx28 route on average during a weekday?

Answer: 14,000

(b) Below is a histogram of the weight (in lbs) of 40 cats.



(i) How many cats could weigh more than 14lbs? Give the largest possible number.

Answer: 4, assuming the cat in the bin including 14lbs is actually larger than 14lbs.

(ii) Could any of the cats weight exactly 8lbs? No explanation of answer needed.

Answer: No.

(c) Assume the parking ticket dataset (see last page) has been loaded into the dataframe `parking` and that all necessary import statements are already included.

Your graphs should have titles and axis labels.

(i) Write a piece of Python code to create a histogram of the data in the `fine_amount` column.

Answer:

```
parking["fine_amount"].hist()
plt.title("Distribution of fine amounts for tickets vehicles")
plt.xlabel("Fine amount in dollars")
plt.ylabel("# of tickets")
```

(ii) Write a piece of Python code to create a bar plot of the data in the `precinct` column.

Answer:

```
counts = parking["precinct"].value_counts()
counts.plot(kind = "bar")
plt.title("Distribution of tickets by precinct")
plt.xlabel("Precinct")
plt.ylabel("# of tickets")
```

3. Assume the parking ticket dataset (see last page) has been loaded into the dataframe `parking` and that all necessary import statements are already included.

(a) Write a piece of Python code to calculate the median year of ticketed vehicles.

Answer:

```
parking["year"].median()
```

(b) Write a piece of Python code that calculate the standard deviation of the fine amount for all tickets given to Ford vehicles.

Answer:

```
ford_filter = parking["vehicle"] == "FORD"
parking[ford_filter]["fine_amount"].std()
```

(c) Write a piece of Python code to calculate the mean year of vehicles given a fine greater than \$50. **Answer:**

```
fine_filter = parking["fine_amount"] > 50
parking[fine_filter]["year"].mean()
```

4. Assume the parking ticket dataset (see last page) has been loaded into the dataframe `parking` and that all necessary import statements are already included.

(a) (i) What is the formula for computing the probability that a ticketed vehicle is registered in New Jersey?

Answer: $\frac{\text{\# of ticketed vehicles registered in New Jersey}}{\text{\# of ticketed vehicles}}$

(ii) Write a piece of Python code that estimates the probability that a ticketed vehicle is registered in New Jersey.

Answer:

```
nj_filter = parking["state"] == "NJ"
num_nj = parking[nj_filter].sum()
num_tickets = parking.shape[0]

num_nj/num_tickets
```

(b) (i) What is the formula for computing the probability that a ticketed vehicle made in 2012 was given a fine of $40 or more?

Answer: $\frac{\text{\# of ticketed vehicles made in 2012 with a fine of \$40 or more}}{\text{\# of ticketed vehicles made in 2012}}$

(ii) Write a piece of Python code that estimates the probability that a ticketed vehicle made in 2012 was given a fine of $40 or more.

Answer:

```
year_filter = parking["year"] == 2012
fine_filter = parking["fine_amount"] >= 40

num_from_2012 = year_filter.sum()

num_large_fine_from_2012 = parking[year_filter & fine_filter].shape[0]

num_large_fine_from_2012/num_from_2012
```

5. (a) The code below uses the `parking` dataframe described on the last page of the exam.

```
mean_list =[]
for i in range(5000):
    sample = parking.sample(100)
    mean_list.append(sample["year"].mean())
pd.Series(mean_list).hist()
```

(i) What is the sample size in the above code?

Answer: 100

(ii) How many samples are simulated in the above code?

Answer: 5000

(iii) What does the `append()` function do in the second last line of code?

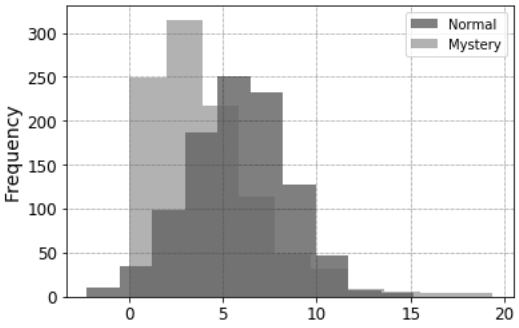
Answer: Add the mean of the years in the sample into the Python list `mean_list`.

(iv) What does the `pd.Series()` function do in the last line of code?

Answer: Convert the Python list my_list into a Pandas series.

(b) The plot below shows two overlapping histograms of 1000 samples from two different distributions. The first sample is drawn from a normal distribution

with a mean of 6 and a standard deviation of 2.5. The second sample is drawn from a mystery distribution.



(i) Is the mean of the mystery distribution smaller or larger than the mean of the normal distribution?

Answer: Smaller

(ii) You are playing a game where you get a single number sampled from each distribution, choose a distribution, and win $100 if the number sampled from your chosen distribution is larger. To have the best chance of winning the money, do you choose the normal or the mystery distribution? Why?

Answer: I choose the normal distribution, because from the plot, we see that its larger values have a higher frequency than the larger values fo the mystery distribution. The mystery distribution has a higher frequency of smaller values.

6. At a certain bookstore, 45% of people who enter the store make a purchase. The bookstore has a sale one week, and during that week 1,008 customers enter the store and 483 of those customers make a purchase. The bookstore owner wants to know if the sale increased the percentage of customers making a purchase.

(a) Which of the following could be the null hypothesis?

(i) The percentage of customers making a purchase remained at 45%.

(ii) The percentage of customers making a purchase increased.

(iii) The percentage of customers making a purchase decreased.

Answer: (i) The percentage of customers making a purchase remained at 45%.

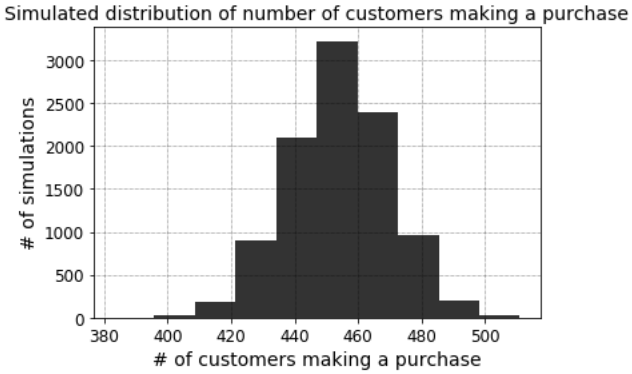
(b) Write a piece of Python code to simulate 10,000 samples of 1,008 customers, where each customer has a 45% chance of making a purchase. The number of customers making a purchase in each sample should be stored in a list.

Answer:

```
population = ["purchase","no purchase"]
pop_prob = [0.45,0.55]

num_purchased_list = []
for i in range(10000):
        sample = np.random.choice(population, p= pop_prob, size = 1008)
        sample_counts = pd.Series(sample).value_counts()
        num_purchased_list.append(sample_counts["purchase"])
```

(c) Suppose the histogram of the number of customers who made a purchase in each of the 10,000 samples is below. Based on this histogram and the data recorded by the bookstore, do you think the sale increased the percentage of customers making a purchase? Why or why not?



Answer: No, there is not enough evidence to reject the null hypothesis and say that the sale increased the percentage of customers making a purchase. From the histogram, there is still a reasonable, if low, probability that 483 customers make a purchase even if the percentage of customers making a purchase hasn't increased.

7. The dataset `solar`, which is also used in Question 9, contains information about solar panel installations in New York city. The columns are:

   - `project_cost` - the cost of the installation project in dollars
   - `incentive` - the amount of incentives in dollars paid by New York state
   - `capacity_rating` - the total capacity of the solar panels in kilowatt hours
   - `expected_kWh` - the expected annual electricity production from the solar panels in kilowatts

   (a) The correlation matrix for the `solar` dataframe is shown below.

	project_cost	incentive	capacity_rating	expected_kWh
project_cost	1.000000	0.816102	0.825615	0.824467
incentive	0.816102	1.000000	0.865710	0.868624
capacity_rating	0.825615	0.865710	1.000000	0.999711
expected_kWh	0.824467	0.868624	0.999711	1.000000

(i) Which two columns are the most correlated?

Answer: capacity_rating and expected_kWh

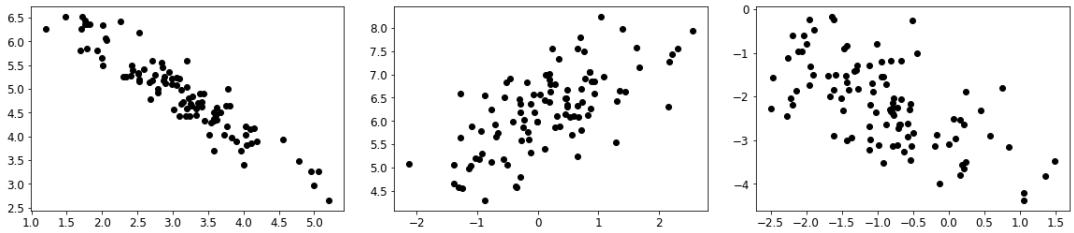
(ii) Which two columns are the least correlated?

Answer: project_cost and incentive

(iii) If you wanted to predict the project cost for a solar panel installation project using only *one* of the other variables, which variable would you choose? Why?

Answer: capacity_rating because this variable is the most correlated with project_cost

(b) The scatter plots below represent data with correlations -0.934, -0.627, and 0.671, in some order. Write the correct correlation below each plot.



Answer: -0.934,0.671,-0.627

8. (a) Suppose the mean delay at a station for a subway train is 6 minutes, with a standard deviation of 3 minutes. A subway improvement lobbyist measures the delay of 75 randomly selected trains, and computes the mean delay. What distribution does this sample mean come from and why? Be as precise as possible.

Answer: By the Central Limit Theorem, the sample mean comes from a normal distribution with a mean of 6 minutes and a standard deviation of $\frac{3}{\sqrt{75}}$

(b) In general, what is the sampling distribution of the mean? You can use a picture in answering this question.

Answer: The sampling distribution of the mean is the distribution of the means of all possible samples taken from the population.

9. The dataset `solar`, which is also used in Question 7, contains information about solar panel installations in New York city. The columns are:

- `project_cost` - the cost of the installation project in dollars
- `incentive` - the amount of incentives in dollars paid by New York state

- `capacity_rating` - the total capacity of the solar panels in kilowatt hours
- `expected_kWh` - the expected annual electricity production from the solar panels in kilowatts

The following Python code creates a linear model from the data in `solar`:

```
lm = smf.ols(formula = 'project_cost ~ incentive + capacity_rating', data = solar).
```

(a) What is/are the dependent variable(s)?

Answer: The project cost.

(b) What is/are the independent variable(s)?

Answer: The amount of incentives and the capacity rating.

(c) Running the code `lm.summary()` gives the following output:

OLS Regression Results

Dep. Variable:	project_cost	**R-squared:**	0.723
Model:	OLS	**Adj. R-squared:**	0.720
Method:	Least Squares	**F-statistic:**	328.3
Date:	Thu, 16 May 2019	**Prob (F-statistic):**	6.64e-71
Time:	12:35:20	**Log-Likelihood:**	-3413.8
No. Observations:	255	**AIC:**	6834.
Df Residuals:	252	**BIC:**	6844.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-627.7654	1.21e+04	-0.052	0.959	-2.45e+04	2.33e+04
incentive	2.4183	0.396	6.104	0.000	1.638	3.199
capacity_rating	1971.0060	274.795	7.173	0.000	1429.819	2512.193

Omnibus:	240.127	**Durbin-Watson:**	1.932
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	32922.327
Skew:	3.020	**Prob(JB):**	0.00
Kurtosis:	58.336	**Cond. No.**	7.42e+04

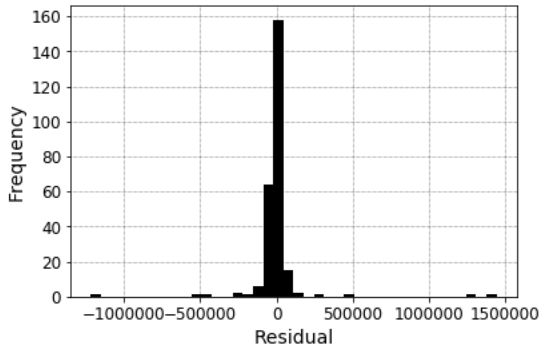
What is the equation of the regression line? Let x_1 represent the amount of incentives and x_2 represent the capacity rating.

Answer: $y = 2.4183x_1 + 1971.0060x_2 - 627.7654$

(d) What is the R-squared value for this model and what does it mean? **Answer:**

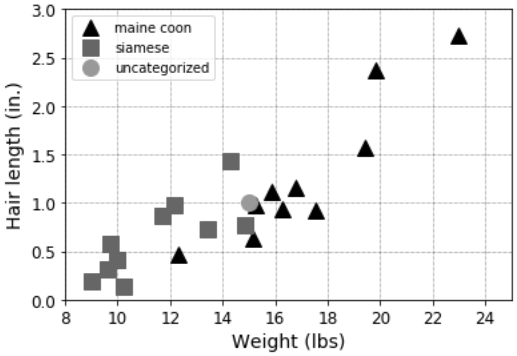
The R-squared value is 0.723 and it represents the variance in the project cost that is explained by the linear model.

(e) A histogram of the residuals is below. Based of this histogram, do you think the linear model is a good fit? Why or why not?



Answer: The histogram of the residuals has several outliers, but otherwise appears normal. Without the outliers, it would be a good fit, but the outliers may not balance out, leading to a model that is too dependent on the outliers.

10. Consider two breeds of cats: Siamese and Maine Coons. Weights (in lbs) and average hair length (in inches) for 10 cats of each breed were recorded and are shown in the scatter plot below. Information about a cat of an unknown breed is also shown (as a circle).



(a) Consider an unclassified cat that weighs 15 lbs and has a hair length of 1 inch, which is represented by a circle in the above scatter plot.

(i) Suppose a 3-nearest neighbor classifier is used. What would be the classification (breed) of this cat? Indicate which data points are used in making the classification.

Answer: "calm", because two of the three nearest neighbors are classified as "calm".

(ii) Suppose a 5-nearest neighbor classifier is used. What would be the classification (breed) of this cat? Indicate which data points are used in making the classification.

Answer: "calm", because three of the 5 nearest neighbors are classified as "calm".

(b) In class, we look at a dataset of passengers on the Titanic that was split into training and testing data. What is the difference between the training and the testing data, and how did we use each one?

Answer: The training data contained a column indicating whether each passenger survived, while the testing data was missing this column. We used the training data to fit the nearest neighbor classifier, and then tested the classifier by classifying the testing data. We then check whether our predictions were correct on Kaggle.

New York Parking Violation Dataset

This dataset contains information about all parking tickets issued in New York City in 2019. Each row represents one parking ticket. Assume that the data has been read in from a csv file and is stored as a Pandas dataframe in the variable `parking`.

summons_num	state	vehicle	precinct	fine_amount	year
347792	NY	HONDA	52	40.0	2014
347793	NJ	FORD	12	40.0	2016
347794	NY	FORD	34	20.0	2008
347795	CT	TOYOTA	35	60.0	2009
347796	NY	SUBARU	78	40.0	2012
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
347834	NY	CHRYSLER	90	20.0	2018

The columns are:

- summons_num = parking ticket number
- state = state vehicle is registered in
- vehicle = vehicle make
- precinct = precinct vehicle was ticketed in
- fine_amount = amount of ticket in dollars
- year = year of vehicle