Sample Final Exam

NAME:

MAT 128, SPRING 2018

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 `traffic.csv` contains information about the number of traffic tickets given out each day in the boroughs of New York. Write a **complete** Python program (excluding import statements) that plots the number of traffic tickets given out in the Bronx over time. Your program should explicitly read in the `Date` column as a date-time object. Your graph should have a title and axes labels.

   `traffic.csv`:

   ```
   Date,Bronx,Brooklyn,Manhattan,Queens,Staten Island
   2019-04-28,65,78,73,83,30
   2019-04-29,69,100,120,82,35
   2019-04-30,75,105,110,81,27
   2019-05-01,83,110,99,90,33
   2019-05-02,77,96,112,85,38
   2019-05-03,80,90,108,43
   2019-05-04,62,85,79,89,32
   ```

   Answer:

   ```
   traffic = pd.read_csv("traffic.csv",parse_dates=["Date"])
   traffic.plot(x = "Date",y="Bronx")
   plt.title("Traffic tickets in the Bronx")
   plt.xlabel("Date")
   plt.ylabel("# of tickets")
   ```

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

   - qualitative or quantitative
   - discrete or continuous
   - nominal, ordinal, interval, or ratio

   (i) Number of teachers at each school in New York City.

   Answer: quantitative,discrete,ratio

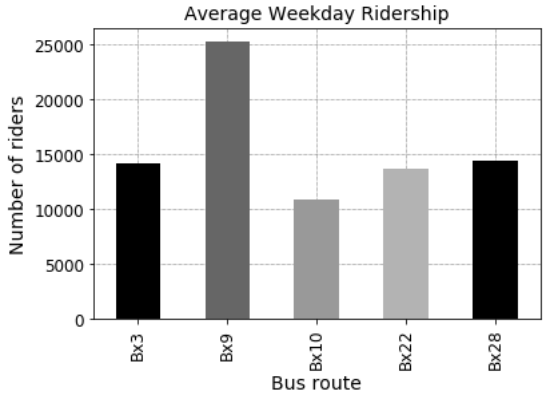
   (ii) Favorite ice cream flavor of 200 randomly selected people visiting Coney Island one afternoon.

   Answer: qualitative, discrete or N/A, nominal

   (iii) Average daily temperature (in Fahrenheit) in each subway station in May.

   Answer: quantitative, continuous, interval

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



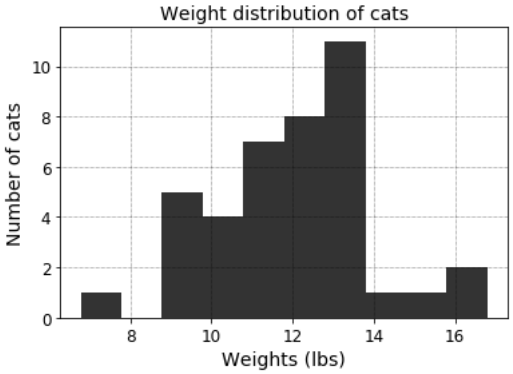
(i) Which bus route has the highest average weekday ridership?

Answer: Bx9

(ii) How many people ride the Bx10 route on average during a weekday?

Answer: 10,500

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



(i) How many cats weigh less than 8lbs?

Answer: 1

(ii) One bin contains more than 10 cat weights. What is the range of weights in this bin?

Answer: 12.8lbs - 13.8lbs

(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 `year` column.

Answer:

```
parking["year"].hist()
plt.title("Distribution of years for tickets vehicles")
plt.xlabel("Vehicle year")
plt.ylabel("# of tickets")
```

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

Answer:

```
counts = parking["state"].value_counts()
counts.plot(kind = "bar")
plt.title("Distribution of vehicle registration state")
plt.xlabel("Registration state")
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 precinct with the most parking tickets.

   Answer:

   ```
   parking["precinct"].mode()
   ```

   (b) Write a piece of Python code to calculate the variance of the fine amounts given to vehicles that are 2010 or older.

   Answer:

   ```
   old_filter = parking["year"] <= 2010
   parking[old_filter]["fine_amount"].var()
   ```

   (c) Write a piece of Python code that calculate the mean fine amount of all tickets given to vehicles from New Jersey (NJ).

   Answer:

   ```
   nj_filter = parking["state"] == "NJ"
   parking[nj_filter]["fine_amount"].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 a Honda?

   Answer: $\frac{\text{\# of ticketed vehicles that are Hondas}}{\text{\# of ticketed vehicles}}$

(ii) Write a piece of Python code that estimates the probability that a ticketed vehicle is a Honda.

Answer:

```
honda_filter = parking["vehicle"] == "HONDA"
num_hondas = parking[honda_filter].sum()
num_tickets = parking.shape[0]

num_hondas/num_tickets
```

(b) (i) What is the formula for computing the probability a vehicle ticketed in the 52 precinct is from 2015?

Answer: $\frac{\text{\# of ticketed vehicles ticketed in the 52nd precinct with year 2015}}{\text{\# of vehicles tickets in the 52nd precinct}}$

(ii) Write a piece of Python code that estimates the probability that a vehicle ticketed in the 52 precinct is from 2015.

Answer:

```
precinct_filter = parking["precinct"] == 52
year_filter = parking["year"] == 2015

num_in_52 = precinct_filter.sum()

num_in_52_from_2015 = parking[precinct_filter & year_filter].shape[0]

num_in_52_from_2015/num_in_52
```

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

```
my_list =[]
for i in range(1000):
    sample = parking.sample(50)
    my_list.append(sample["fine_amount"].std())
pd.Series(my_list).hist()
```

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

Answer: 50

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

Answer: 1000

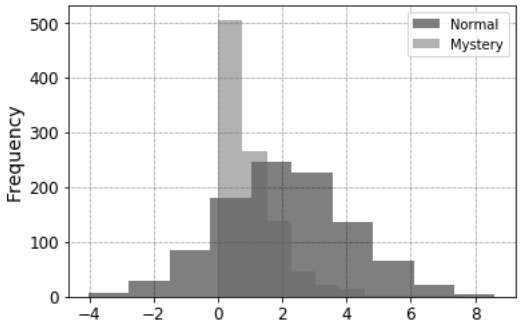
(iii) What is stored in the variable `my_list`?

Answer: The standard deviation of the fine amount of each sample.

(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 2 and a standard deviation of 2. The second sample is drawn from a mystery distribution.



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

Answer: Smaller

(ii) You are playing a game where you get a single number sampled from a distribution, and win $100 if that number is between 0 and 1. To win the money, do you want this number to be sampled from the normal distribution or the mystery distribution? Why?

Answer: I want this number to be sampled from the normal mystery distribution, because more of its distribution is between 0 and 1 than in the normal distribution. Therefore, it is more likely a number randomly selected from the mystery distribution will be between 0 and 1 than a number randomly selected from this normal distribution.

6. A city claims it repairs 90% of potholes within 10 days of the pothole being reported. A reporter looks at data from 2018, and finds that in that year,124 potholes were reported and 102 of the potholes were repaired within 10 days. The reporter believes the city may not be repairing as many potholes within 10 days as it claims, and is going to test this hypothesis.

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

(i) In 2018, more than 90% of the reported potholes were repaired within 10 days.

(ii) In 2018, 90% of reported potholes were repaired within 10 days.

(iii) In 2018, less than 90% of reported potholes were repaired within 10 days.

Answer: (ii) In 2018, 90% of reported potholes were repaired within 10 days.

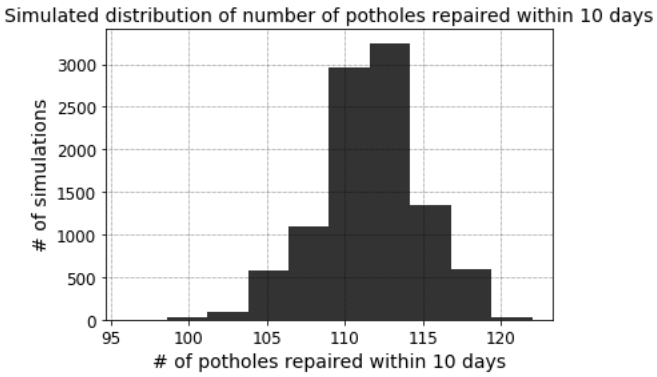
(b) Write a piece of Python code to simulate 10,000 samples of 124 potholes, where each pothole has a 90% chance of being repaired within 10 days. The number of potholes repaired within 10 days in each sample should be stored in a list.

Answer:

```
population = ["repaired","not repaired"]
pop_prob = [0.9,0.1]

num_repaired_list = []
for i in range(10000):
        sample = np.random.choice(population, p= pop_prob, size = 124)
        sample_counts = pd.Series(sample).value_counts()
        num_repaired_list.append(sample_counts["repaired"])
```

(c) Suppose the histogram of the number of potholes repaired within 10 days in each of the 10,000 samples is below. Based on this histogram and the pothole data from 2018, do you think the city is accurate in its claim that it repairs 90% of all reported potholes within 10 days? Why or why not?



Answer: No, the city seems to be over-estimating the number of potholes repaired within 10 days. The data point of 102 repaired potholes is not very likely, and on the lower end, of the distribution of repaired potholes if the city really was repairing 90% of the potholes within 10 days. Therefore, it seems likely that the city is not correct in its claim.

7. The dataset `times`, which is also used in Question 9, contains information about buildings in Times Square. The columns are:

   - `property_type` - the type of building (retail, office, residential, etc)
   - `num_elevators` - the number of elevators in the building
   - `year_built` - the year the building was constructed
   - `land_area` - the size of the property in acres
   - `rentable_area` - the rentable area in the building in square feet

- `num_elevators` - the number of elevators in the building. `num_stories` - the number of stories in the building

(a) The correlation matrix for the columns `num_elevators`, `year_built`, `land_area`, `rentable_area`, and `num_elevators` in the `times` dataframe is shown below.

	year_built	num_stories	land_area	rentable_area	num_elevators
year_built	1.000000	0.638714	0.543932	0.591299	0.491487
num_stories	0.638714	1.000000	0.704419	0.862790	0.821792
land_area	0.543932	0.704419	1.000000	0.815492	0.723483
rentable_area	0.591299	0.862790	0.815492	1.000000	0.880626
num_elevators	0.491487	0.821792	0.723483	0.880626	1.000000

(i) Which two columns are the most correlated?

Answer: num_elevators and rentable_area

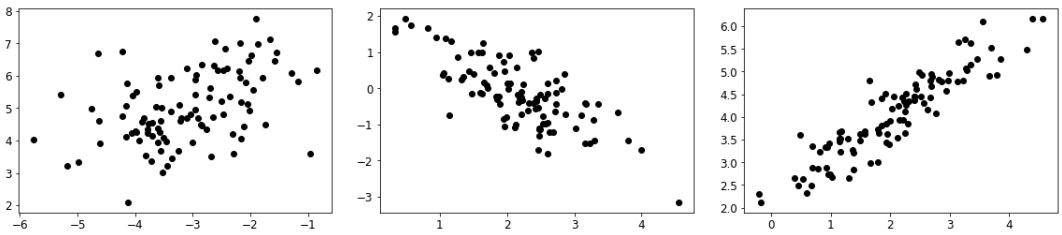
(ii) Which two columns are the least correlated?

Answer: year_built and num_elevators

(iii) If you wanted to predict the number of elevators in a building using only *one* of the other variables, which variable would you choose? Why?

Answer: rentable_area because this variable is the most correlated with num_elevators

(b) The scatter plots below represent data with correlations -0.753, 0.442, and 0.918, in some order. Write the correct correlation below each plot.



Answer: 0.442,-0.753, 0.918

8. (a) Suppose the mean number of steps walked each day by a New Yorker is 7,340 steps with a standard deviation of 528 steps. A researcher has 50 randomly selected New Yorker track their steps for a day, and computes the mean number of steps taken. 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 7,340 steps and a standard deviation of $\frac{528}{\sqrt{50}}$

(b) 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 `times` contains information about buildings in Times Square. The columns are:

- `property_type` - the type of building (retail, office, residential, etc)
- `num_elevators` - the number of elevators in the building
- `year_built` - the year the building was constructed
- `land_area` - the size of the property in acres
- `rentable_area` - the rentable area in the building in square feet
- `num_elevators` - the number of elevators in the building. `num_stories` - the number of stories in the building

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

```
lm = smf.ols(formula = 'num_elevators ~ num_stories + rentable_area', data = times)
```

(a) What is this model predicting?

Answer: The number of elevators in a building.

(b) What information is this model using to make the prediction?

Answer: The number of stories in the stories in the building and the rentable area of the building square feet.

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

Dep. Variable:	num_elevators	**R-squared:**	0.790
Model:	OLS	**Adj. R-squared:**	0.788
Method:	Least Squares	**F-statistic:**	299.7
Date:	Thu, 09 May 2019	**Prob (F-statistic):**	1.14e-54
Time:	11:43:44	**Log-Likelihood:**	-426.35
No. Observations:	162	**AIC:**	858.7
Df Residuals:	159	**BIC:**	868.0
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.5207	0.380	-1.369	0.173	-1.272	0.231
num_stories	0.1305	0.039	3.384	0.001	0.054	0.207
rentable_area	1.097e-05	1.17e-06	9.340	0.000	8.65e-06	1.33e-05

Omnibus:	205.499	**Durbin-Watson:**	2.308
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	10927.416
Skew:	-4.938	**Prob(JB):**	0.00
Kurtosis:	42.005	**Cond. No.**	7.01e+05

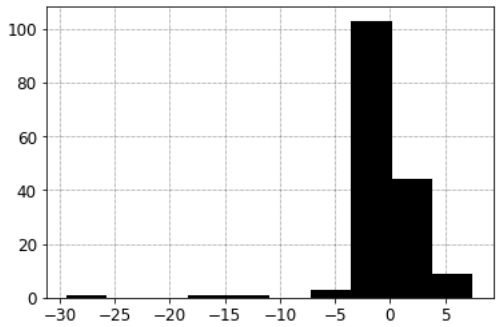
What is the equation of the regression line? Let x_1 represent the number of stories and x_2 represent the rentable area.

Answer: $y = 0.1305x_1 + 0.00001097x_2$

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

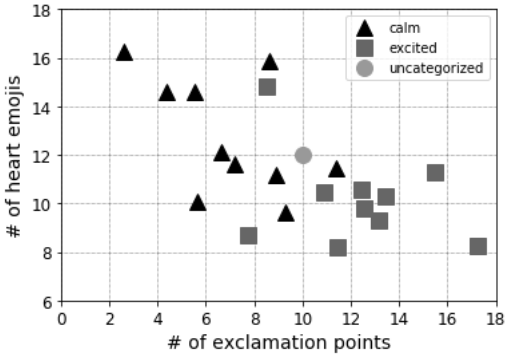
The R-squared value is 0.79 and it represents the variance in the number of elevators explained by the linear model.

(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 somewhat normal. The linear model would be a better fit without the outliers, but is probably a good enough fit with them.

10. A social media company wants to classify photos as "exciting" or "calm" based on the number of exclamation points and the number of heart emojis in the comments for each photo. The company has a training set of 20 "calm" photos and 20 "exciting" photos, along with the number of exclamation point and heart emojis in the comments of each photo. This training data is shown in the scatter plot below.



(a) Consider an unclassified photo with 10 exclamation points and 12 heart emojis in the comments, 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 of this photo?

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 of this photo?

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 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