CSE 351 HW 1

1.

a.

https://www.wego.com/schedules has a database of publicly available flight schedules. It's possible to check what flights are available, at whatever time necessary and to wherever necessary. It is public since there is no restricted access for internet users.

b.

https://www.ivywise.com/blog/college-admission-rates/ has a list of universities and their admission statistics. It is possible to see acceptance rate and when, clicking on the hyperlinks, we can also see more detailed description of admissions such as specific year enrollment rate and graduate enrollment. This is publically accessible data, not behind any requirements.

c.

Using https://compstat.nypdonline.org/ we can determine the type of crime, place and time of the crime in the new york city area. This is accessible to the public as it can be easily searched up.

d.

IMDB pro offers a way to check box office revenue for movies. However this is locked behind a paywall thus it could be considered private data instead of public data

e.

The site https://www.wunderground.com/history/daily/us/ny/new-york-city gives access to daily temperature data in the new york city area. It is detailed to the hour and also has data from previous years. The data here is publically accessible and can be accessed through a quick google search.

2. The public dataset called Electric Power Monthly, by the US Energy information administration provides monthly information on US energy usage. This is done by what type of resources is used to produce the energy, as well as who the user of the energy is. Three things that this can be used for is, understanding solar power energy generation trends, total energy usage trends in the US, and finding a correlation between coal usage and solar panel usage.

Another public dataset called "Baby Names from Social Security Card Applications - National Data" from US Government's open data shows the name, year of birth and

sex of social security card applicants from 1880 onward. Using this data it is possible to understand naming trends throughout the last 145 years. We can also use this data to find out what the most common names for each sex is. Along with that it is possible to also find some correlation between last names and first names.

Finally we also have the public dataset "Real Estate Sales 2001-2022 GL". This is also from US Government's open data. It provides information on real estate sales between the years 2001-2022 between October 1-September 30. The data is from the state of Connecticut. It includes, town, address, date of sale, type of property, sales price, and assessment of property. This type of data can be used to create an automatic price calculator for properties in the area. It can also be used to understand if property values are rising within the area. We can also create a map of the area based on what type of properties are most valuable in that general location.

3.

Randomly select 10 people from your friend group. Pour diet coke and regular coke each into an enclosed bottle so that your friends cannot see the liquid. Label the regular coke 1 and the diet coke 2. Have your friends each do a blind taste test of both bottles and write down which they prefer onto a sheet, 1 or 2. Analyze the data.

4.

Logarithms can be used while working with probabilities. Instead of multiplying probabilities, which can lead to really small numbers, we can take the log of the probabilities and add them together. Another use is for when a graph is exponential. It is possible to make the scale logarithmic and this can lead to a linear plot which is easier to look at and fit. Logarithms can also be used to make more sense of ratios. Using log, it is possible to get the same magnitude value when ratios are the same. For example $log_2(1/2) = -1$ and $log_2(2) = 1$.

5.

Correlation is when two values have a noticeable trend between them. One of the values can be used to make a somewhat accurate prediction of the other value. Causation on the other hand is when one event leads to another. For example, rainy weather causes the ground to be wet. While correlation could imply causation, it is not always the case. For example the number of drownings and ice cream sales could be correlated due to the weather, but that doesn't mean ice cream sales cause drowning.

6.

a.

The mean of X is 8.125 hours. The mean of Y is 73.9.

b.

For X the standard deviation is 3.5684. For Y the standard deviation is 13.3334

c.

The covariance between X and Y is 47.303.

d.

The pearson correlation coefficient is $r = 0.9942$. This indicates a strong positive linear correlation

7.

a.

the rank of X is (5 10 9 12 15 4 13 2 17 7 8 14 1 19 11 20 6 3 18 16) and the rank of Y is (5 10 8 11 15 4 13 2 18 7 9 14 1 19 12 20 6 3 17 16)

b.

The difference between the ranks $rank(x_i) - rank(y_i)$ is (0 0 1 1 0 0 0 0 -1 0 -1 0 0 0 -1 0 0 0 1 0)

c.

The sum of the squared difference is 6

d.

The spearman rank correlation coefficient, ρ is then

$$1 - (6 * 6/(20(20^2 - 1))) = 0.9955$$

which indicates a strong monotonic positive relation.

8.

The pearson and spearman correlation both give values in the 0.99 range. This is because of the strong monotonic positive relationship between X and Y. The similarities between the two coefficients is that they both range from 1 to -1, where positive indicates a positive correlation and negative indicates a negative correlation. Pearson coefficient explains that the square of r which is $r^2 = 0.9942^2 = 0.9884$ indicates

that 98.84% of the variance in Y can be explained by a linear fit. The spearman correlation indicates that there are some points that don't have matching X and Y ranks. However the majority do and the ones that don't are not significantly off in the Y and X ranks, thus strongly monotonic.

9.

a.

ii has a greater mean and standard deviation

b.

i has the greater mean, however ii has the greater standard deviation

c.

ii has the greater mean but they both have the same standard deviation

d.

They both have the same mean however ii has a larger standard deviation

10.

a.

No because we don't know if A and B are independent events.

b.

i.

We have $P(A \; and \; B) = P(A) * P(B) = 0.21$

ii.

We have $P(A \; or \; B) = P(A) + P(B) - P(A)P(B) = 0.79$

iii

We have $P(A|B) = P(A \cap B)/P(B) = P(A)P(B)/P(B) = 0.3$

11

Probability is the chances of an event occurring and predicting the chances of future events occuring. This can be with or without prior knowledge. Statistics is analyzing real world data for trends and patterns. It is used to understand correlation between multiple events in the real world.

12

We have $P(D+) = 100/2000 = 5\%$. We have $P(T+) = 200/2000 = 10\%$. We have $P(D+|T+) = 95/200 = 47.5\%$. We have $P(T+|D+) = 95/100 = 95\%$. We have $P(T-|D-) = 1795/1900 = 94.47\%$. The test seems to be 95% effective. This is because when someone does have a disease it can detect that disease 95% of the time. Though the test may have false positives, these aren't life threatening.