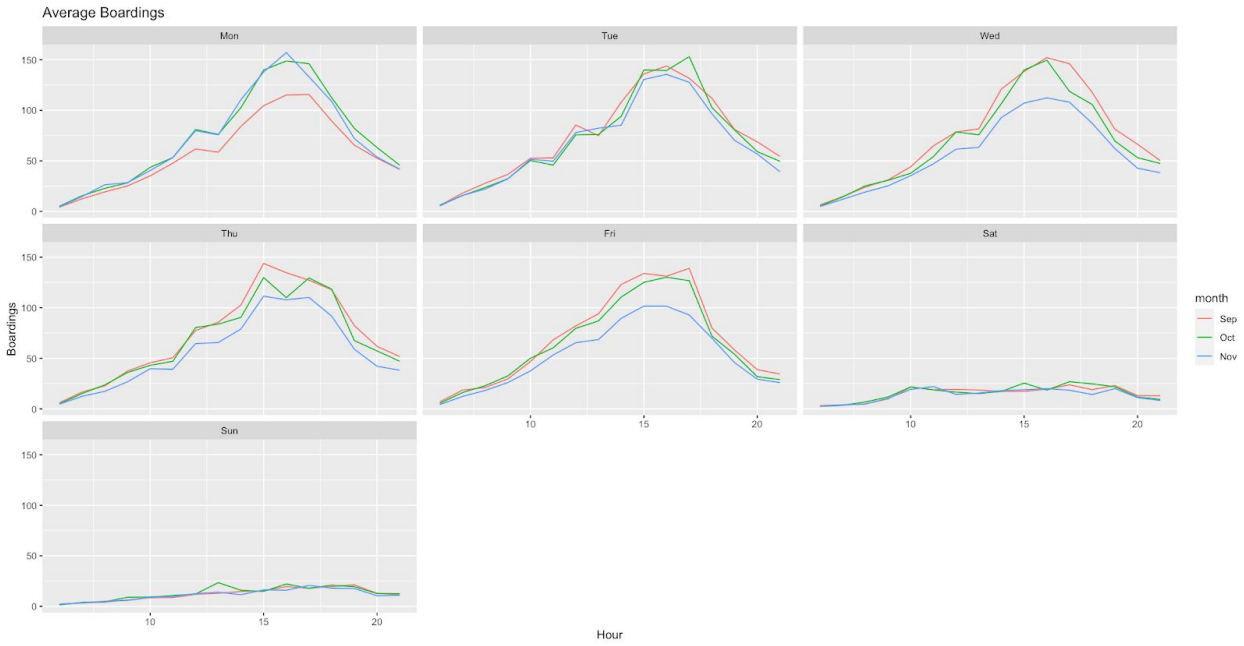
Homework 5

STA 309

Problem 1

Problem 1.A

Figure 1



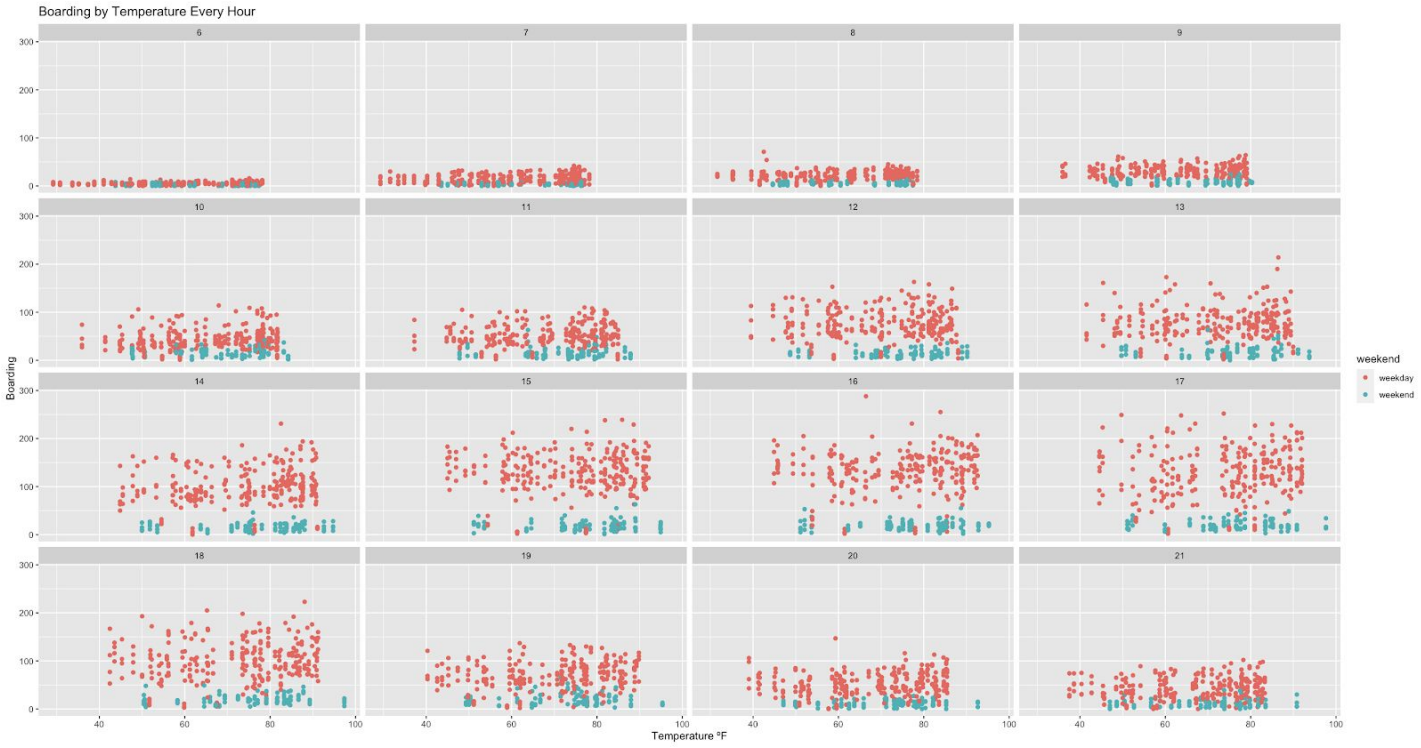
Line plots faceted by the day of the week, showing the hour (x) by the average number of bookings (y), colored by the month.

Caption

- Looking at the graph, it can be seen that the hour of peek boardings stay about the same for weekdays, but change on weekends, as seen by the last two facets in the graph.
- On average, it seems that average bookings on Monday in September are lower than that of October and March, per the color coded months in the monday facet.
- I believe that the number of bookings, potentially, on Tuesday, Wednesday and Thursday in November are lower on average due to the fact that November is closer to the holidays so there may be averaging lower fares in addition to the change in temperature as November is near-winter.

Problem 1.B

Figure 2



The number of boardings (y) by temperature (x) for each 15 minute time interval, color coded by whether or not it is a weekday (with red being weekday), and faceted by hour.

Caption

1. Holding everything else constant, temperature does not seem to have a large impact on the number of boardings by UT students as with each marginal temperature increase, the number of bookings does not seem to change by much (i.e. there does not appear to be a relationship between bookings and temperature). Looking at the graph, when temperature increases, it appears to be an almost identical amount to those when there is a temperature decrease.

Problem 2.A

Question

Run a convincing analysis that shows that green certification is statistically significant (i.e. confidence intervals do not contain 0)

Approach

- For the outcome metric, I decided to measure "success" by using Revenue per Square foot because not all of the rent will be translated to revenue, because it is a direct metric to the amount of revenue that will be generated. For example, an apartment complex may charge green-based premiums, like EV charging.
- For the green metric, I decided to use the Green Rating variable, as it takes into account the LEED rating and Green Star ratings as well.
- For the methodological approach, I will build a multiple regression model, by regressing the model 1,000 times under a Monte Carlo simulation and determining confidence intervals.
- For the confounders, I used green_rating, age, class, and City_Market_Rent, as those are some of the most relevant variables to rent amount, as opposed to the others such as number of stories in the building.

Results

Figure 1

	Lower	Upper	Estimate
Intercept	-1.81	0.68	-0.56
Green Rating	0.70	2.10	1.42
Age	-0.01	0.01	0.00
Class B	-4.95	-3.86	-4.41
Class C	-9.17	-7.51	-8.40
City_Market_Rent	0.96	1.05	1.00

A table detailing the confidence intervals reported for each of the corresponding confounders,

LEED, age, class, and City Market Rent, in attempts to predict the Rent variable.

Conclusion

It appears that the Green Rating is one that is statistically significant, due to the fact that 0 does not appear within the confidence intervals at a 95% level. The revenue appears to increase between 0.070 to 2.10 dollars per square foot, estimated to be 0.647 per square foot; this is, the partial relationship between rent and green rating. Age, Class and City Market Rent are all parts of the relationships as it is multiple regression.

Problem 2.B

Question

Run a convincing analysis that shows that green certification is not statistically significant (i.e. confidence intervals do contain 0)

Approach

- For the outcome metric, I decided to measure "success" by using Revenue per Square foot because not all of the rent will be translated to revenue, because it is a direct metric to the amount of revenue that will be generated. For example, an apartment complex may charge green-based premiums, like EV charging.
- For the green metric, I decided to use the Green Rating variable, as it takes into account the LEED rating and Green Star ratings as well.
- For the methodological approach, I will build a multiple regression model, by regressing the model 1,000 times under a Monte Carlo simulation and determining confidence intervals.
- For the confounders, I used green_rating, age, class, and City_Market_Rent, as those are some of the most relevant variables to rent amount, as opposed to the others such as number of stories in the building; I also included the interaction between age and Green rating as it is more likely that newer buildings will be greener.

Results

Figure 2

	Lower	Upper	Estimate
Intercept	-1.8279	0.8995	-0.4707
Green Rating	-1.9131	0.9341	-0.1853
Age	-0.0072	0.0127	0.0012
Class B	-4.8902	-3.9109	-4.4290
Class C	-9.0718	-7.6908	-8.3850
City Market Rent	0.9576	1.0466	1.0025
Green Rating : Age	0.0142	0.1612	0.0656

A table detailing the confidence intervals reported for each of the corresponding confounders,

LEED, age, class, and City Market Rent and Green Rating : Age, in attempts to predict the Revenue Per Square Feet variable.

Conclusion

It appears that the Green Rating Variable (even when taking into account the interaction between Green Rating and Age) is statistically insignificant, due to the fact that the confidence interval ranges between -1.91 to 0.93 of an increase in revenue per square foot. This is a partial relationship between age, class, city market rent, and the interaction between green rating and age.

Problem 2C

Table

Freedoms	Part A	Part B	Winner
Outcome	I used the Revenue by SQFT outcome as not all of the revenue is strictly rent (e.g. Green premiums)	Same as Part A	Revenue by SQFT
Green	I used Green Rating as it incorporates two measures of green, rather than just one	Same as Part A	Green Rating
Approach	Multiple regression over a Monte Carlo Simulation (1,000) and determining confidence intervals	Same as Part A	Multiple Regression
Confounders	Green Rating, Age, Class, City Market Rent. Most direct impacts on revenue and rent.	Green Rating, Age, Class, City Market Rent, and Green Rating:Age. Younger buildings tend to be geared more green.	Part B, because it takes into account the age interaction to the building, while having a partial relationship with the same confounders.

Conclusion

The conclusion of this p-hacking experiment would be that the Green Rating is a statistically insignificant predictor of success on the commercial real estate market as the analysis with Part B, where an interaction between Green Rating and Age was included with all else being the same in Part A, provides a confidence interval that goes through 0 -- all else being the same outcome predictor, green predictor, approach, and same confounders aside from the interaction. Thus it should be concluded that the green rating is statistically insignificant to the analysis of commercial real estate.

Problem 3

Question

Fit separate exponential growth models for Italy and Spain using days_since_first_death as the time variable.

- Characterize the doubling time for each country's death total. Are these doubling times similar, or noticeably different from one another?
- Include the confidence intervals for the doubling times, a line graph showing deaths over time, faceted by country.

Approach

To determine the doubling times, I fit an exponential model using the natural log of the number of deaths (y) to the number of days since the first death (x), over 1,000 Monte Carlo Simulations. Once completed, I determined the confidence intervals for the model as well as calculated the doubling times at each respective interval.

Results

Figure 1

ITALY	Lower	Upper	Estimate	Low Double	Upper Double	Estimated Double
Intercept	0.539	1.635	1.019	1.287	0.424	0.680
days_since_first_death	0.158	0.212	0.183	4.379	3.276	3.783

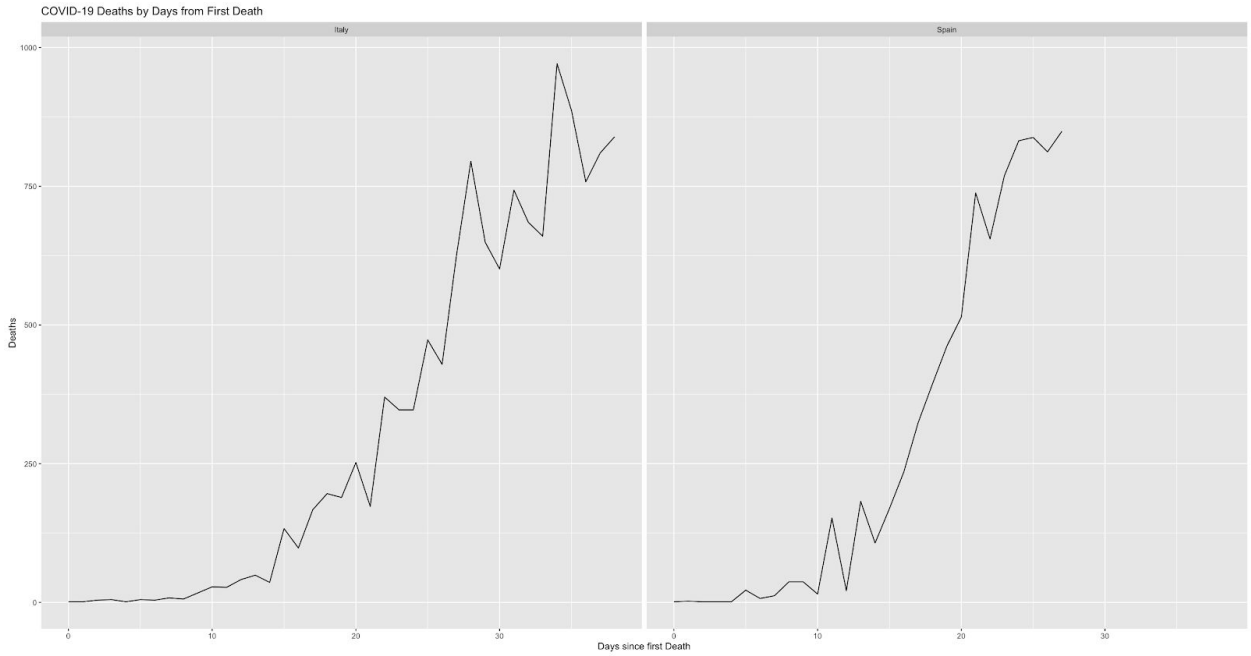
This table shows the confidence intervals for the fit exponential model for Italy, with the corresponding estimates for the doubling time.

Figure 2

SPAIN	Lower	Upper	Estimate	Low Double	Upper Double	Estimated Double
Intercept	-0.133	1.233	0.465	-5.225	0.562	1.490
days_since_first_death	0.235	0.315	0.276	2.950	2.198	2.509

This table shows the confidence intervals for the fit exponential model for Spain, with the corresponding estimates for the doubling time.

Figure 3



This line graph shows the number of deaths (y) for each day after the first death (x), faceted by the country (Italy on the left and Spain on the right).

Conclusion

Recalling the Exponential Growth Model:

$$Y_t = Y_0 \bullet e^{rt}$$

In this case, Y_t is expressed as the number of deaths, Y_0 is the offset amount, r is the rate of exponential change, and t is the number of days since the first death.

This model can be re-expressed on a logarithmic scale

$$log(Y_t) = log(Y_0) + r \bullet t$$

To calculate the doubling time, we can use the r variable.

$$\frac{log(2)}{r}$$

Italy

We know that for Italy, on a log scale, that r for the number of days since the first death is between 0.158 and 0.212, estimated to be about 0.183. The doubling time, on a logarithmic scale, is between 4.379 and 3.276, estimated to be 3.783. This means that for approximately every 3.5-4 days, there was a doubling of the number of deaths due to COVID-19 in Italy, at 95% confidence.

Spain

We know that for Spain, on a log scale, that r for the number of days since the first death is between 0.235 and 0.315, estimated to be about 0.276. The doubling time, on a logarithmic scale, is between 2.950 and 2.198, estimated to be 2.509. This means that for approximately every 2.5 days, there was a doubling of the number of deaths due to COVID-19 in Spain, at 95% confidence.

What it all Means

These doubling times are close to one another (3.5 vs 2.5 days), however, looking at what doubling time means -- the number of days it takes for the deaths to double -- the one day difference makes a big difference; thus, these two doubling times are noticeable different from one another. By analyzing the graph, we see this to be evident as well.