Homework 2 SDS315

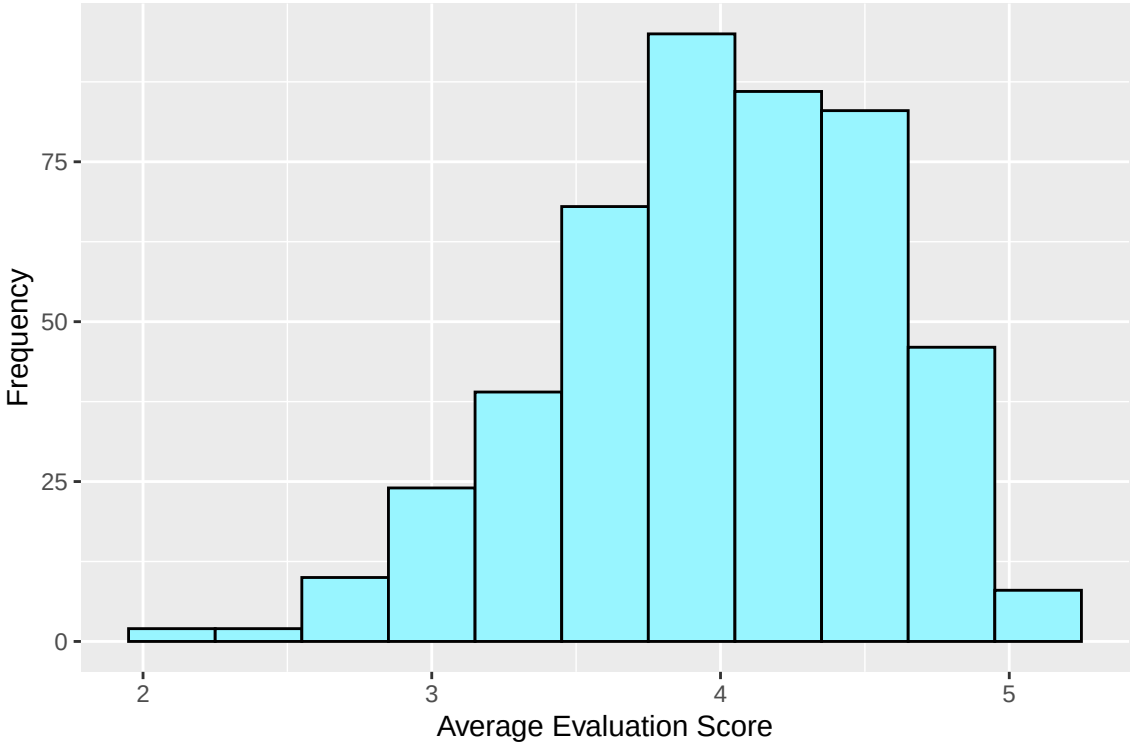
Lucas Chiang (lmc4866) UT Austin

Github Link: https://github.com/leichiang1/sds_315.git

Problem 1 Beauty, or not, in the classroom

Part A. Create a histogram to display the overall data distribution

Average Course Evaluation Scores for Instructors at UT Austin

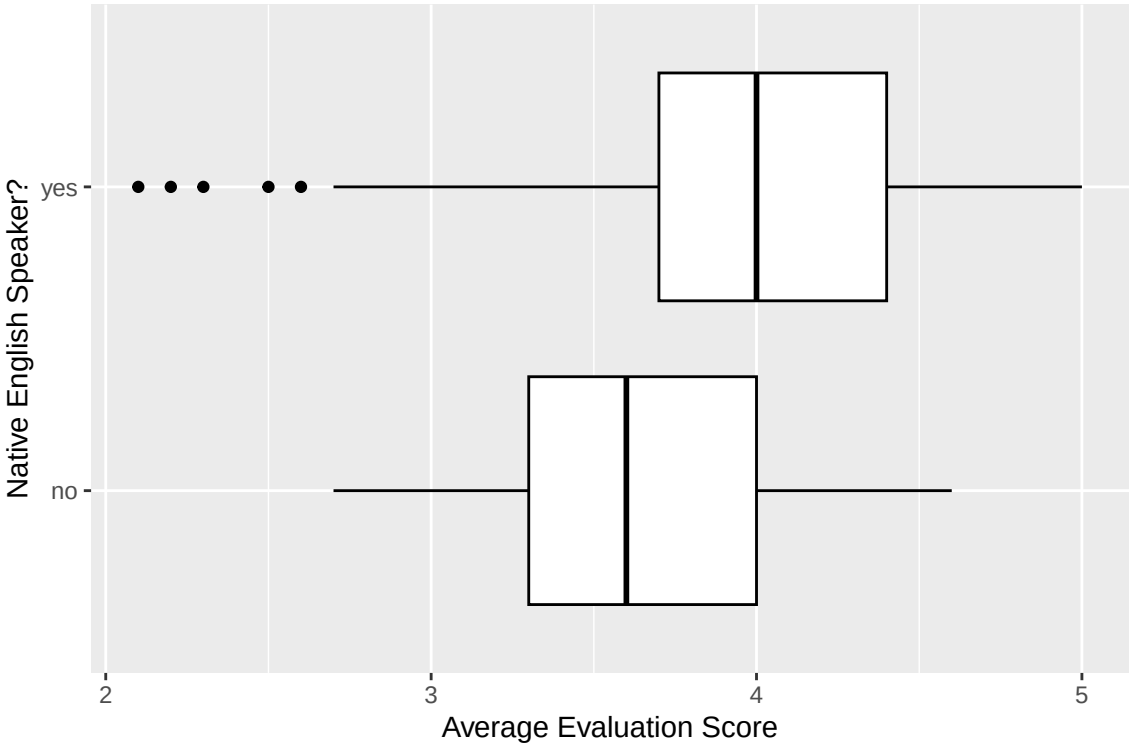


Key Features: Above shows a histogram displaying 463 average course evaluation scores (scale from 1 to 5) of instructors at UT Austin by 25,547 students. Instructors were rated on a scale from 1 to 5.

Takeaways: The distribution of the course evaluation has a slightly left skewed distribution. A majority of scores fall in the 4-5 range, with the mean evaluation score being around 4. This could indicate that students at UT rate their professors rather highly. However, we must take into consideration that 25,547 students does not represent the student population at UT.

Part B. Use side-by-side box plots to show the distribution of course evaluation scores by whether or not the professor is a native English speaker.

Course Evaluation Score Distributions Based on English Speaking Background

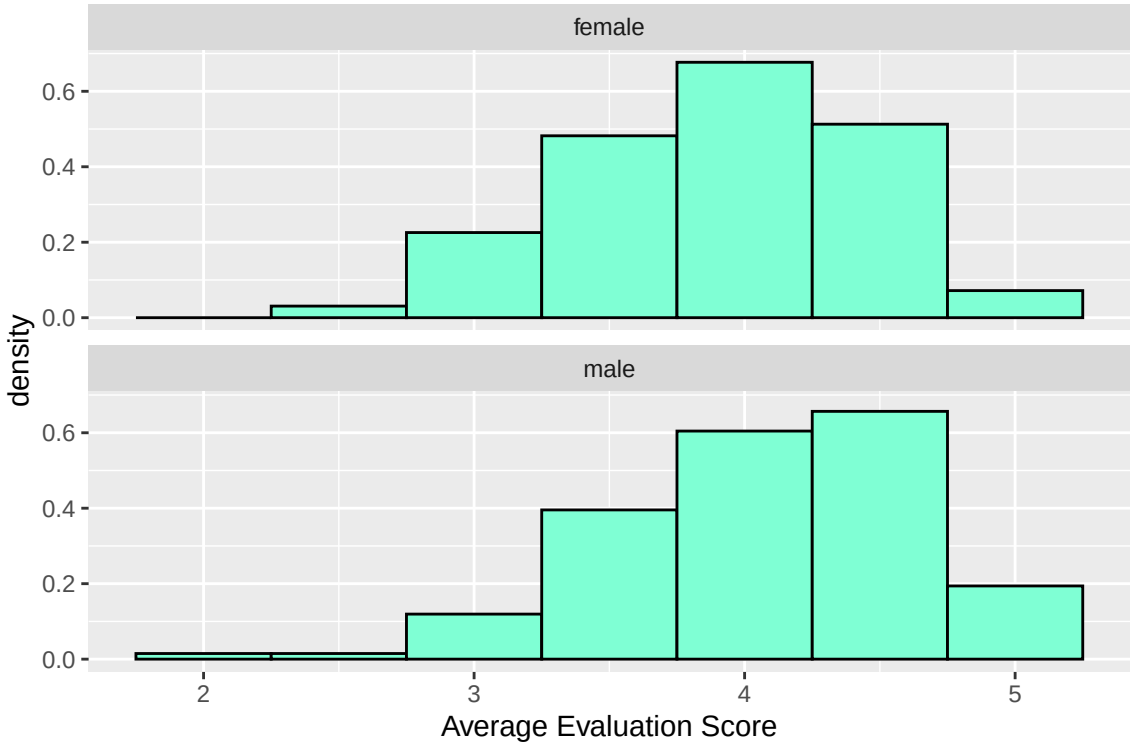


Key Features: Above is the same data of the 463 average course evaluation scores, but this time I have used box plots to represent the distributions. This time, the distributions are divided based on an instructor's English speaking background (whether they are a native speaker or not).

Takeaways: Notice how the instructors that are native English speakers have a higher median average course evaluation score than their non-native English speaking counterparts (indicated by the bold line). Even though this could indicate that non-English speakers are poorer instructors than native speakers, there are a number of factors to consider. First, is that the variation in both distributions is greater than the difference between the averages. Also, a confounding variable could be the difficulty of the classes. Maybe more non-native English speaking instructors are teaching harder classes. Difficulty of classes could be a significant factor in the lower course evaluations without necessarily being in regard to an instructor's English background. Furthermore, it is vague to evaluate distribution like this with instructors being native English speakers or not. A non-native speaker could be fluent in English and not distinguishable from a native speaker. Due to this, not being a native English speaker may not have significant effect on the communication between students and their instructor and thus may not be a contributor to lower course evaluations. The main takeaway is that this plot only gives us a vague idea of the difference between an native English speaking instructor and one who is not.

Part C. Use a faceted histogram with two rows to compare the distribution of course evaluation scores for male and female instructors.

Course Evaluation Distributions Based on an Instructor's Gender

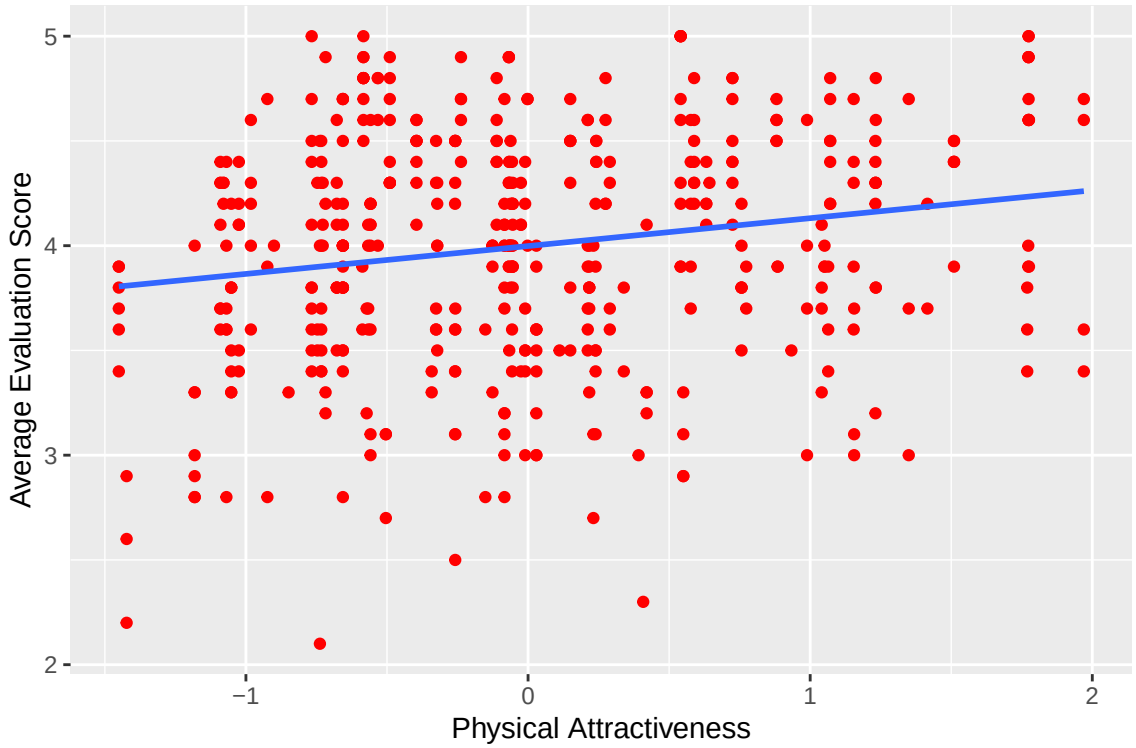


Key Features: The histogram shows the distribution of average course evaluations scores for male and female instructors at UT Austin.

Takeaways: Here we are comparing the course evaluations between male and female instructors. Both distributions of male and female instructors have left skewed distributions since more ratings fall towards the higher end. This is a density histogram meaning that the area under the histogram is one. Visually speaking, it seems like a larger percentage of course evaluation scores for male professors are in the upper 4-5 range than female professors. This is confirmed by the median value 4.15 for male professors, and 3.9 for female professors. This distinction could mean that UT students tend to rate male professors more highly, but remember that this data is not representative of the whole UT student population.

Part D. Create a scatter plot to visualize the extent to which there may be an association between the professor's physical attractiveness (x) and their course evaluations (y).

Scatterplot of Physical Attractiveness vs. Course Evaluation Scores



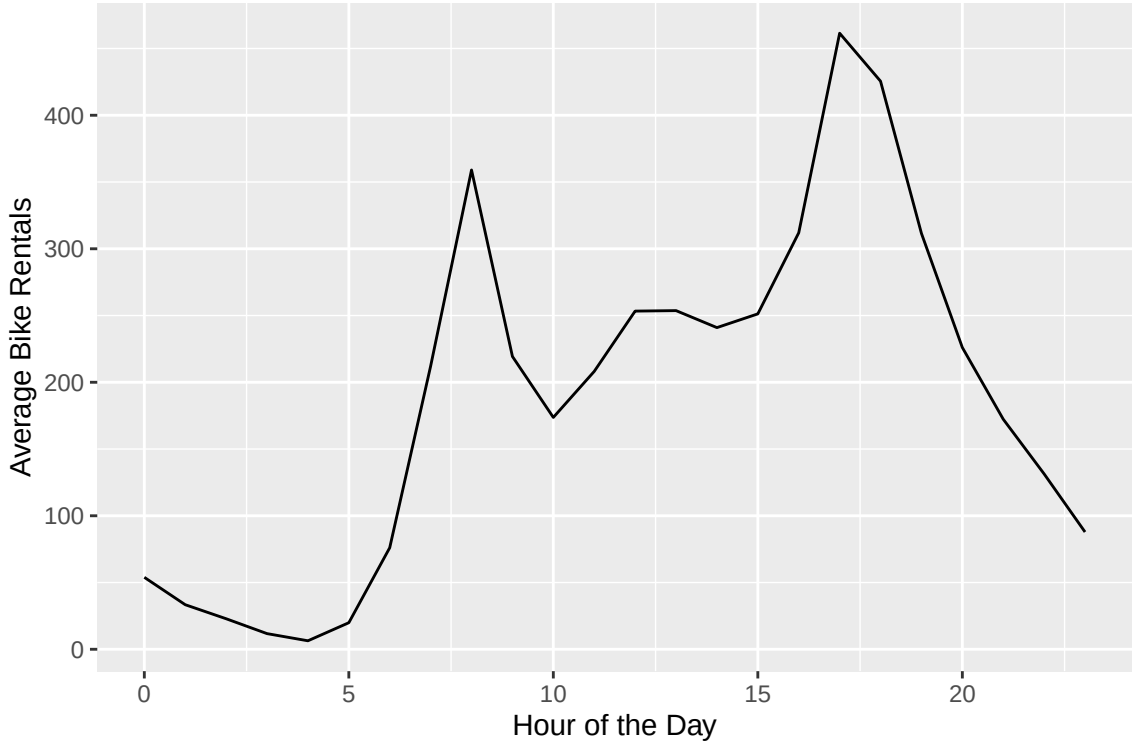
Key Features: The scatter plot above displays the relationship between an instructor's physical attractiveness and their course evaluation scores. Physical attractiveness was determined by a panel of six students and scores are based on points above or below average attractiveness (indicated by 0).

Takeaways: The line of best fit is elevated towards the right hand side which indicates a positive relationship between physical attractiveness and course evaluation scores. However, many points on this scatter plot stray away from the line of best fit which indicates that there is a weak relationship between an instructor's physical outlook and their evaluation scores by students. This is confirmed by the r-value (a measure of correlation) which is .189. An r-value of 1 would indicate a strong, positive correlation between numerical variables, but the r-value here is far from that.

Problem 2: bike sharing

Part A. a line graph showing average hourly bike rentals (total) across all hours of the day (hr).

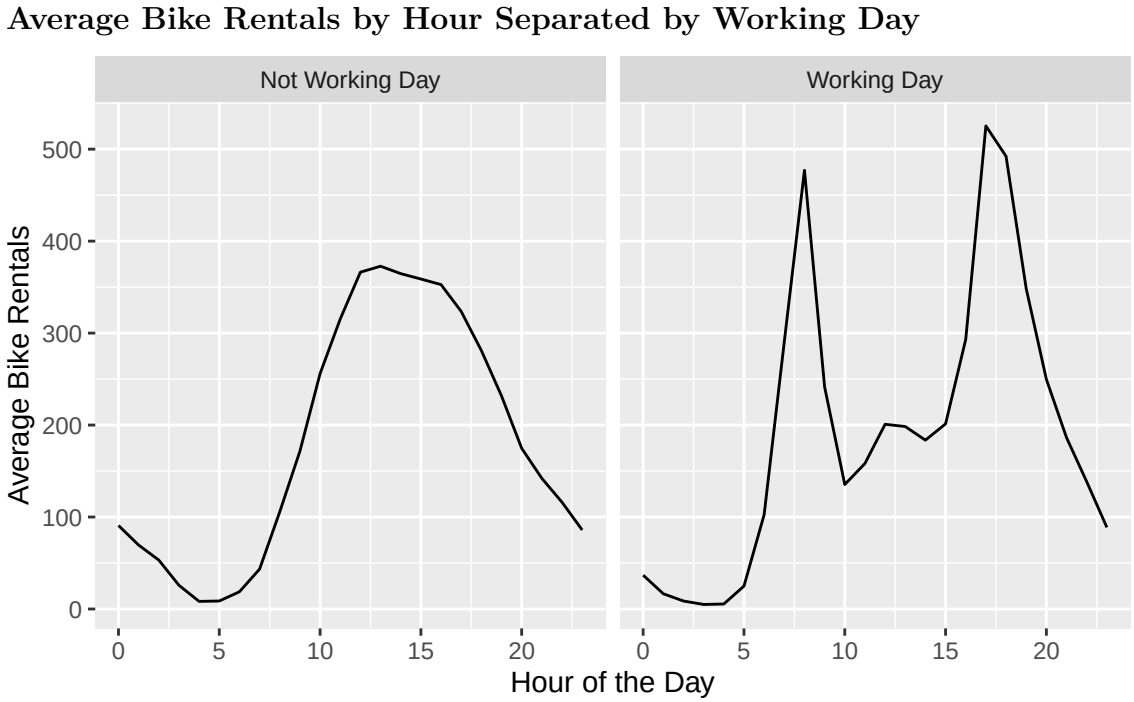
Line Graph for Average Bike Rentals for each Hour in a Day



The line graph shown is a line graph depicting the average bike rentals by the hour from a historical log of bike share data in Washington D.C from 2011 to 2012. As you can see, the x-axis (the lower axis) represents the hour of the day, and the y-axis (on the side) depicts the average number of bike rentals recorded during that particular hour in D.C.. The line graph here clearly gives us two peaks of bike rentals. One peak is between 8AM and 10AM and the highest peak is between 4PM and 5PM. Also notice how there are still quite high large of average bike rentals past 5PM into the evening before eventually falling out. There are a few possibilities for this. The hours of peaks are during the hours where people commute to and commute back from work. This may show that many people in D.C. may use bikes to commute to and from work. People may also use bike rentals as a form of leisure in the afternoon.

Main Takeaway: Bike rentals are highest during the times where people commute to and from work.

Part B. A faceted line graph showing average bike rentals by hour of the day, faceted according to whether it is a working day (workingday)

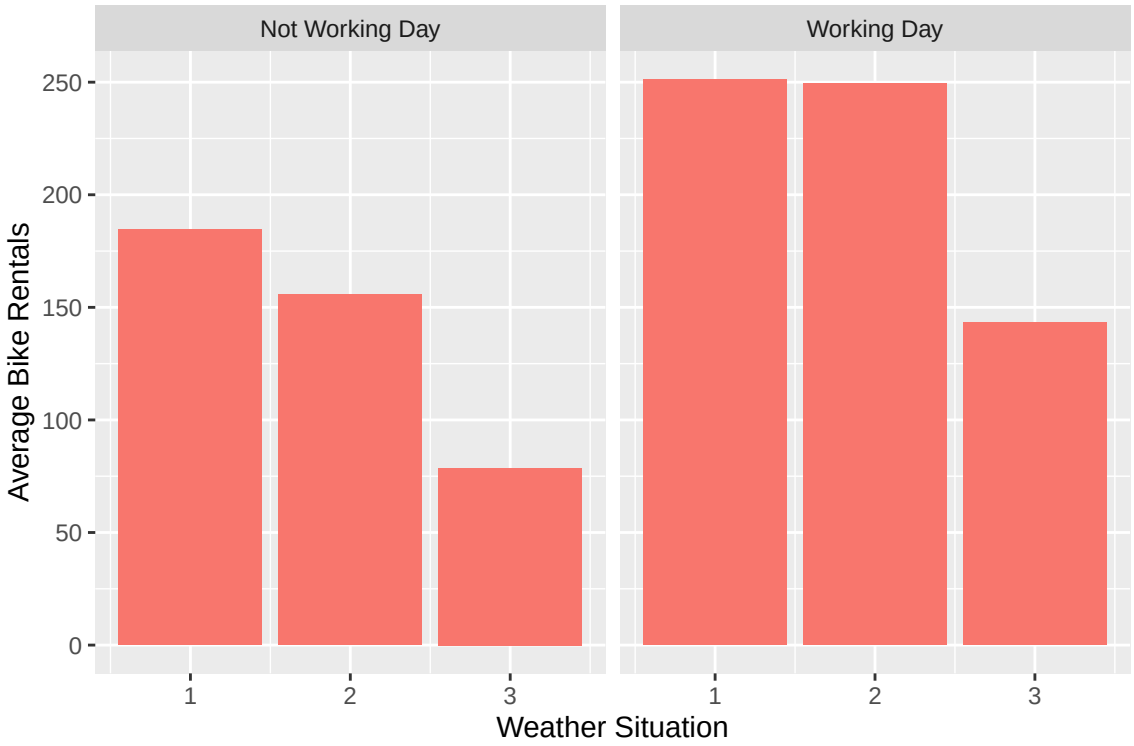


Here is the same bike rental data from Washington D.C., but here I have separated the data and plotted a line graph based on if the day was a working day or not. Again, the x-axis represents the hour of the day and the y-axis indicates the average number bike rentals. Clearly, the two graphs are dissimilar. There are no sharp peaks on the line graph that is not a working day, but a fairly large amount of bike rentals can be seen during the midday from hours 12AM-1PM. On the other hand, the working day line graph maintains the attributes from the combined line graph, having two peaks during the usual time when people commute to and from work. These dissimilarities possibly point towards a conclusion that commuting to and from work creates a sharp demand in bike rentals.

Main Takeaway: Bike rentals have the sharpest peaks in demand during working days, but do not have such peaks during non-working days.

Part C. a faceted bar plot showing average ridership (y) during the 9 AM hour by weather situation, faceted according to whether it is a working day or not.

9 AM Average Bike Rentals Based on Weather Situation



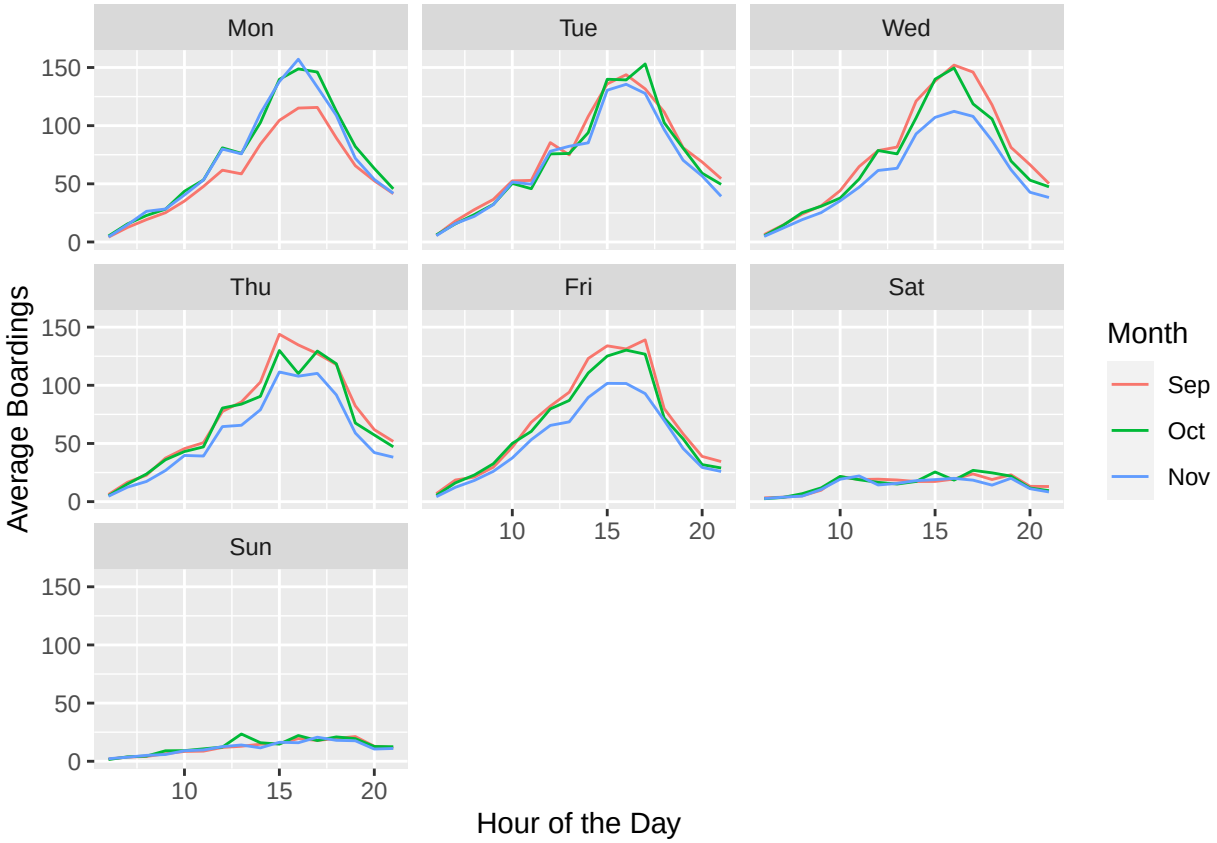
Above is a bar graph showing the average bike rentals in the 9AM hour based on non-working days and working days. Each bar represents a different weather situation that occurred during that time. "1" describes weather that is either clear, few clouds, or partly cloudy. "2" describes whether that is generally misty and cloudy or misty with few clouds. "3" represents weather with either light snow, light rain, or thunderstorms with scattered clouds. There is also a 4th category that involves weather that is more extreme, but none of the data in this particular subset was categorized as such. As we can see in the bar graphs above, the difference between weather situation 1 and 2 in both bar plots is interesting because the difference is greater in non-working days than working days. Slightly worse weather did not have much of a difference in average bike rentals for working days at 9AM likely because many people are commuting to work, which is an important matter for most people. However, we see a steep drop for weather situation 3 in working days. This could probably be explained by people finding other forms of transportation because of precipitation, or maybe work was canceled for the day. However, these are only speculations. In general though, the weather situation does seem to have a correlation with drop in average bike rentals.

Main Takeaway: Worse weather situations correlate with less average bike rentals for both working days and non working days at 9AM, but non-working days have a more gradual drop due to worse weather conditions and working days have a sudden drop when weather involves precipitation.

Problem 3 Capital Metro UT Ridership

Part 1: One facted line graph that plots average boardings by hour of the day, day of week, and month.

Average Boardings for Each Day of the Week by the Hour



Above is a line graph that plots the hour of the day (0-24) and the average boardings of the UT Capital Metro bus network that would occur according to that hour of the day. Each panel represents a particular day of the week and each panel includes three lines for the months September, October, and November.

Does the hour of peak boardings change from day to day, or is it broadly similar across days?

The hour of peak boardings is broadly similar across the days of the week since the faceted line graph shows each day of the week having boarding peaks around 3 to 5pm in the afternoons. This is particularly prevalent on the week days. Sunday doesn't really follow this trend, however, with it's highest peak happening around 12pm-1pm.

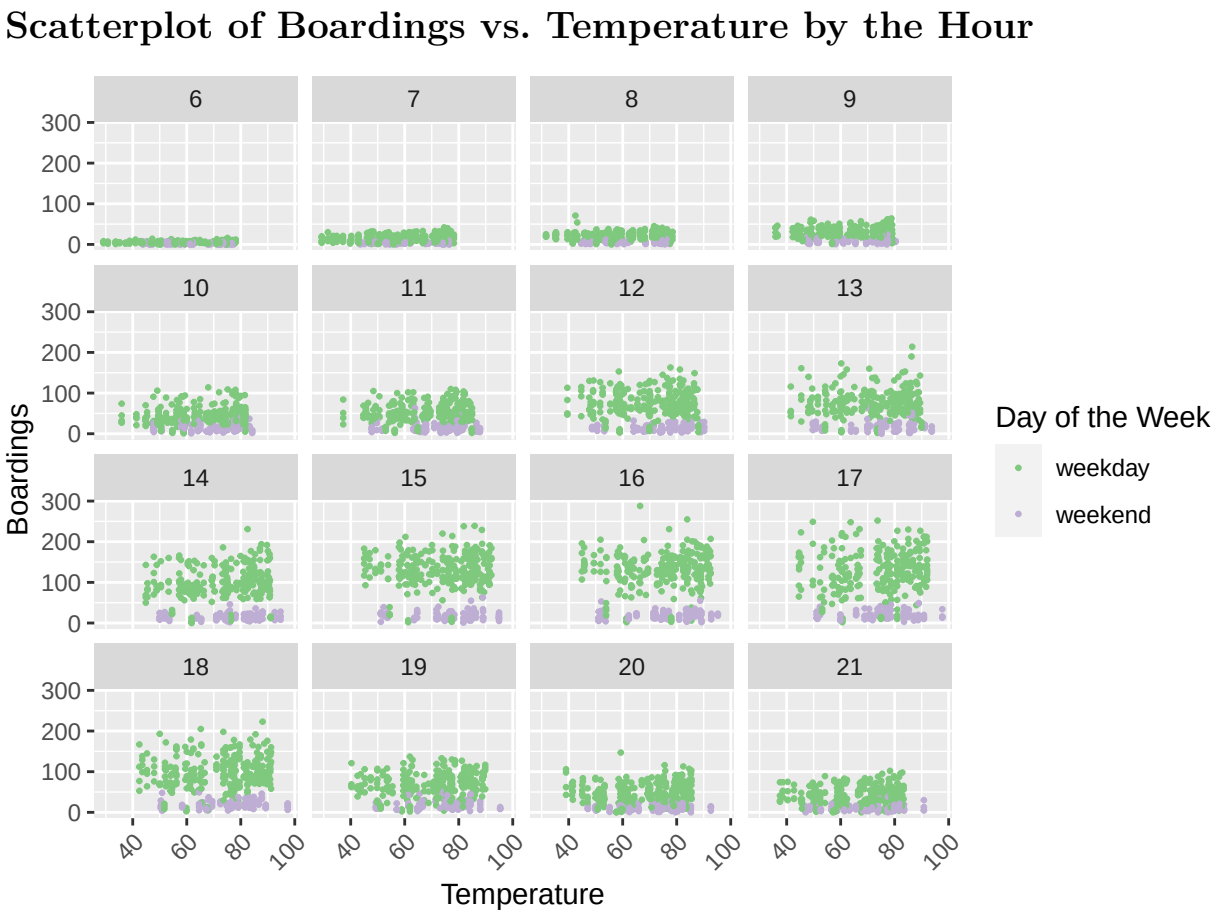
Why do you think average boardings on Mondays in September look lower, compared to other days and months?

I think average boardings on Mondays in September look lower possibly because of the labor day holiday that occurs on the first Monday in September. Across the other days of the week you can observe that there are higher peaks around the time that people get off of work and class (around 5 pm). On the other hand, Monday has a smaller peak of average boardings possibly because a large portion people didn't have work or class on the labor day holiday. This may have plateaued the usual boarding peak for Monday that you would observe as being higher in other days of the week.

Similarly, why do you think average boardings on Weds/Thurs/Fri in Novermber look lower?

I believe the cause of the lower boarding average is possibly very similar to that of the Monday in September. Thanksgiving break that often occurs around Wednesday throughout Friday may have brought down the average numbers. Again we can observe that the peaks have "flattened" during the times that people get off of work and likely end class (around 5pm - 6pm here). Many people, particularly students, may have had a break from work and class from a week on Wednesday-Friday because of Thanksgiving break. This is a possibility of why the average numbers for those days may be lower than the other days of the week.

Part 2: One facetted scatter plot showing boardings (y) vs. temperature (x), faceted by hour of the day, and with points colored in according to whether it is a weekday or weekend.



Above is a scatter plot that plots the number of boardings of the UT Capital Metro to the temperature of that day faceted by the hour of the day. The different colors of points correspond to whether that hour on that particular day was a weekday or weekend. A weekday is represented by green, and a weekend is represented by purple.

When we hold hour of day and weekend status constant, does temperature seem to have a noticeable effect on the number of UT students riding the bus?

No. If we just observe one specific hour on the scatter plot such as the 16th hour, there seems to be to be no upward or downward projection of the number of boardings that occur during that hour regardless of the temperature. The data for the weekend days and the 16th hour points seem to gravitate around the bottom of the plot and stays relatively constant. This is also shown in all of the other hours as well.

Problem 4: Wrangling the Billboard Top 100

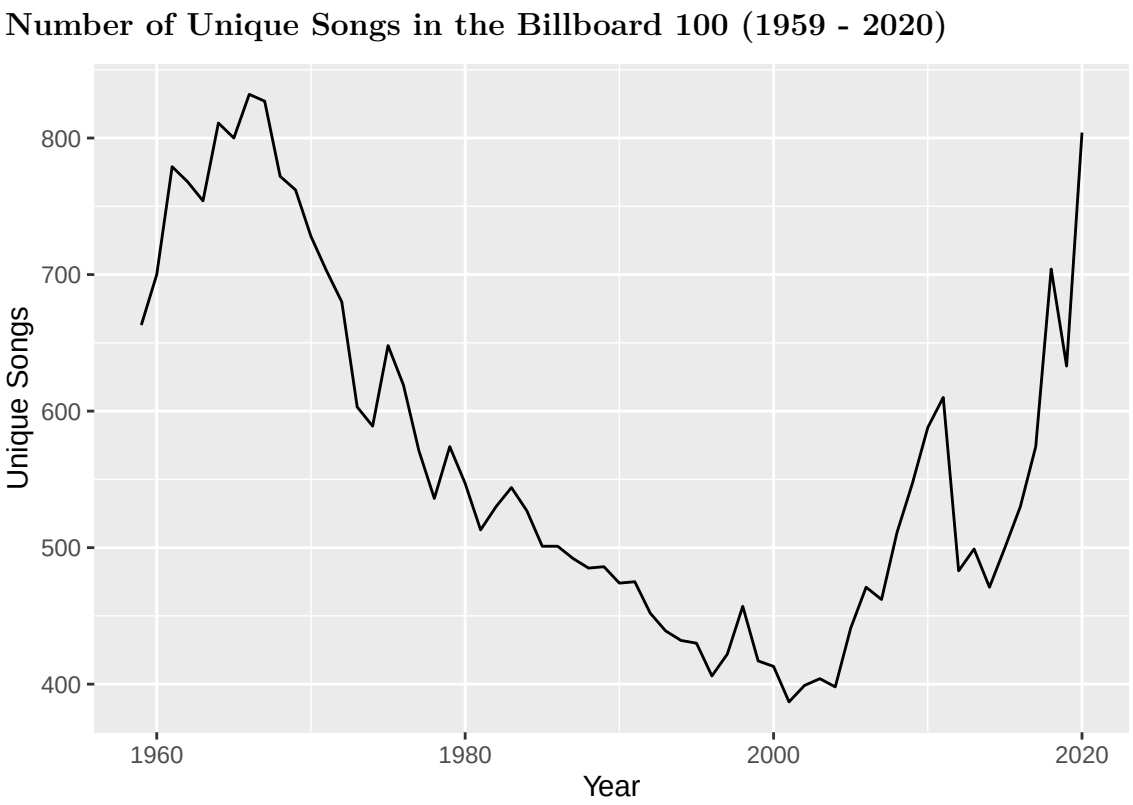
Part A. Make a table of the top 10 most popular songs since 1958, as measured by the total number of weeks that a song spent on the Billboard Top 100.

Most Popular Songs Since 1958

Performer	Song	Weeks in Top 100
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79
Jason Mraz	I'm Yours	76
The Weeknd	Blinding Lights	76
LeAnn Rimes	How Do I Live	69
LMFAO Featuring Lauren Bennett & GoonRock	Party Rock Anthem	68
OneRepublic	Counting Stars	68
Adele	Rolling In The Deep	65
Jewel	Foolish Games/You Were Meant For Me	65
Carrie Underwood	Before He Cheats	64

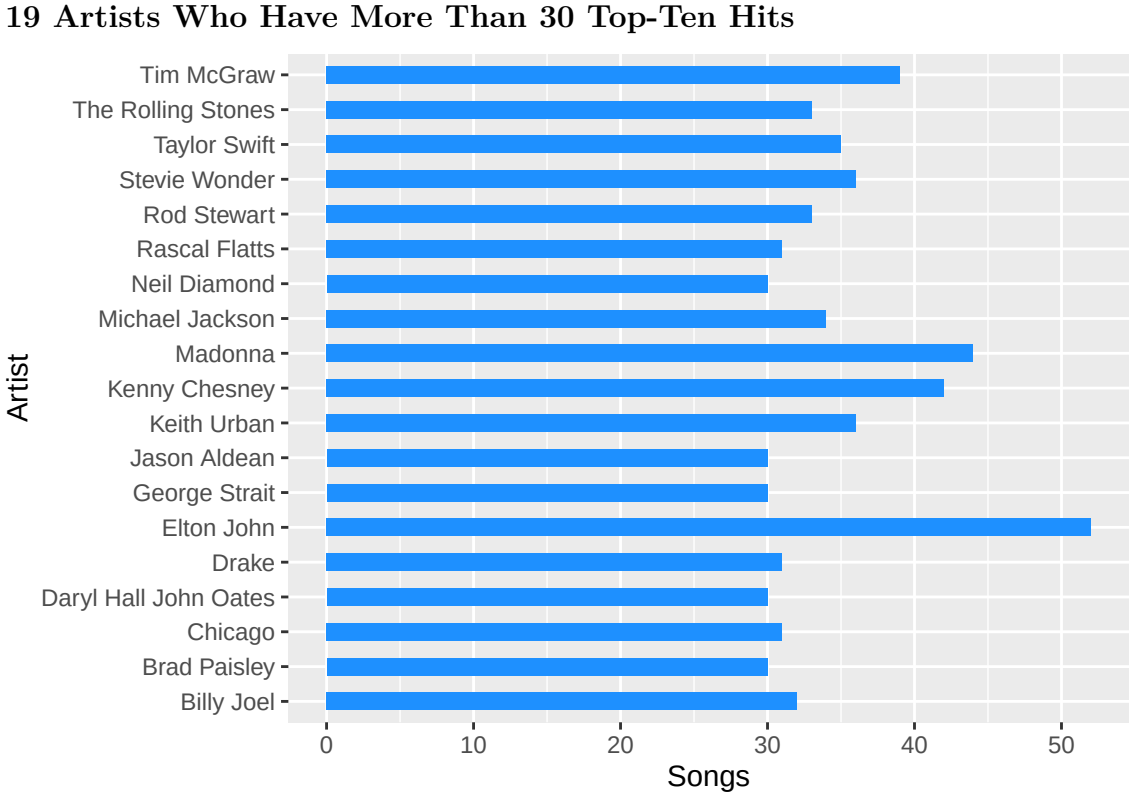
The table above shows the most popular songs according to the Billboard 100 since 1958. The "most popular" songs in this table are determined by the number of weeks that a particular song has appeared in Billboard 100 since 1958.

Part B. Make a line graph that plots the measure of musical diversity over the years.



The line graph displays the number of unique songs that appeared in the Billboard 100 for each year from 1959 - 2020. 1966 was the year that there were the most number of unique songs with a total of 832 songs. An interesting trend happens after this where the number of unique songs declines within the span from 1966-2001. 2001 is where the number of unique songs is the lowest with only 387. However, the uniqueness of songs generally rapidly climbs up again in the next 20 years.

Part C. There are 19 artists in U.S. musical history since 1958 who havef had at least 30 songs that were "ten-week hits." Make a bar plot for these 19 artists, showing how many ten-week hits each one had in their musical career.



Above is a bar plot comparing 19 artists who have at least 30 songs that have been in the Billboard 100 for at least ten weeks. These songs are called "ten-week hits." By observing this bar plot, we can see that Elton John is the artist who had the most ten week hits from 1958 - 2021.