Homework 2

Arun Mahadevan Sathia Narayanan

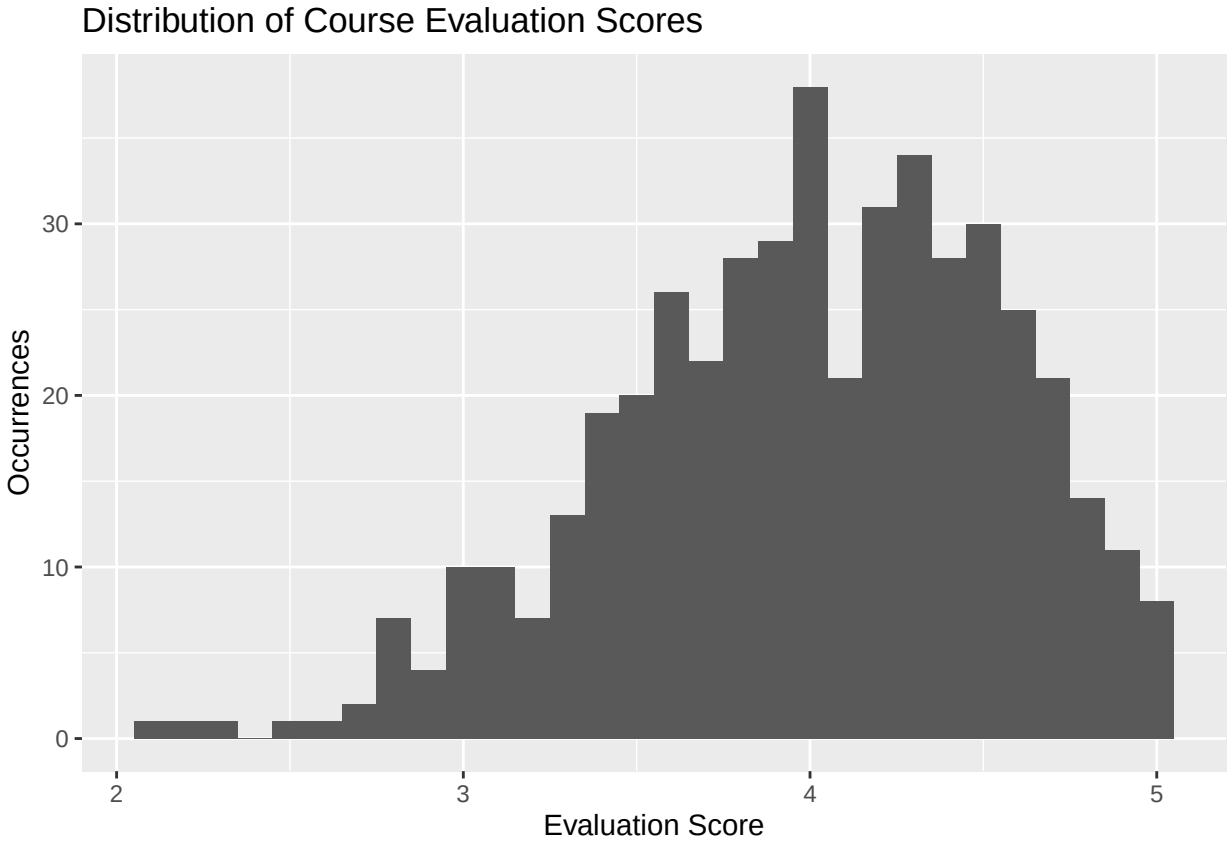
2025-01-26

GitHub Link:
To GitHub

GitHub Link (Text Format):
https://github.com/arunmsn/SDS315/tree/main/Week_03/Homework_2

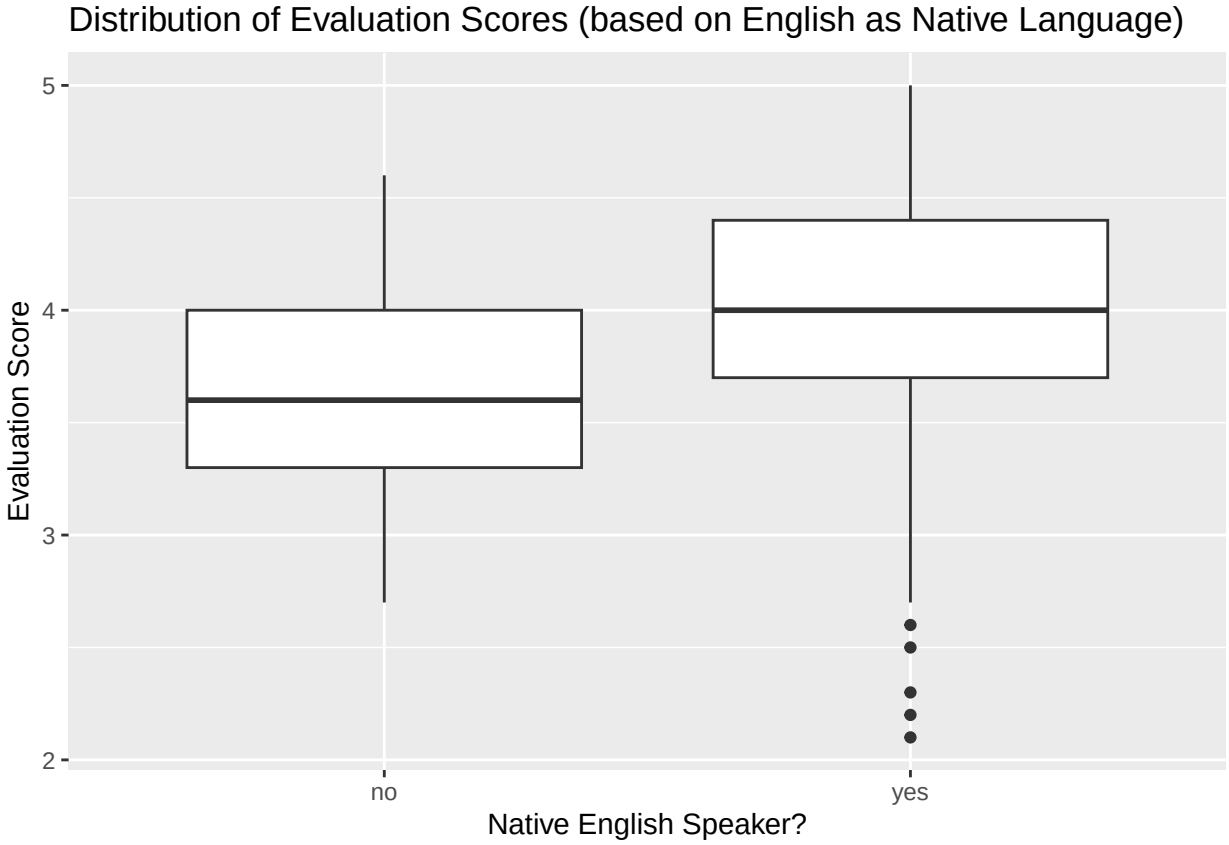
Problem 1 - Beauty, or not, in the classroom

1 - Part A



We can observe that the cluster of the evaluation scores lies near the `3 to 4` range. `4`, in fact, has the highest number of occurrences within the dataset. There are not many courses with very low ratings, with the first cluster of the lower scores showing up past `3`.

1 - Part B



From these side-by-side boxplots we can observe the differences between how the evaluation scores vary based on if the professor's native language is English or not. Immediately observable is the range of values covered by the boxplots. The professors who have English as their native language have more variability in their score than the professors who do not. Let's use numbers rather than visuals to see the data more clearly.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.100   3.700   4.000   4.018   4.400   5.000
```

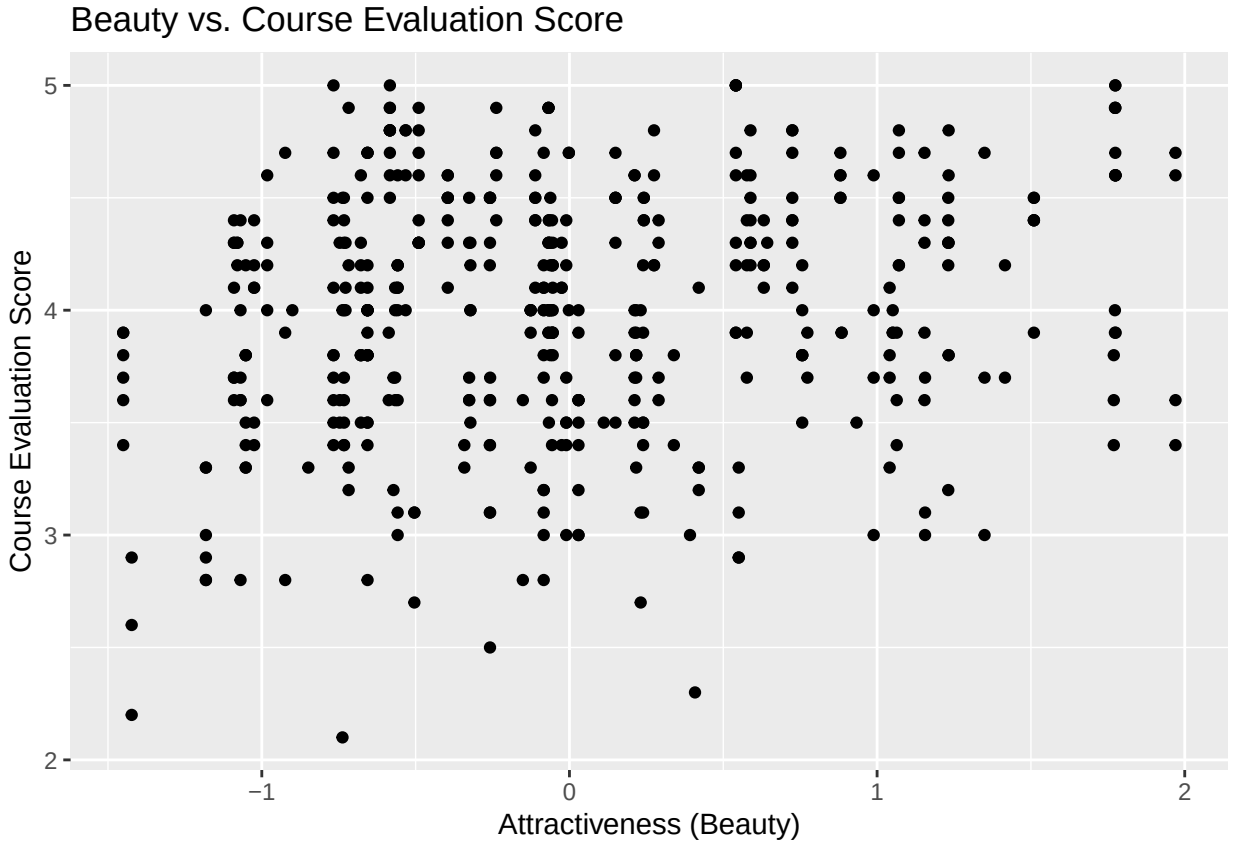
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.700   3.300   3.600   3.689   4.000   4.600
```

The 2 rows on top show the summary for the professors with a native language of English, and the 2 rows on the bottom show the summary for the professors whose native language is not English. From the numbers we can see that the maximum evaluation score for the native-English professors (5.000) is greater than the maximum evaluation score for the non-native-English professors (4.600). On the flip side, the native-English professors have a lower minimum evaluation score compared to the non-native-English professors. Also seen from these numbers is that the average evaluation score and median evaluation score are both higher for the native-English professors compared to the non-native-English professors.

1 - Part C

Using the above, we can tell the difference in evaluation scores based on the professors' genders. When comparing overall, there are more occurrences for male-professor course evaluations compared to female professors. For the female professors, there are more course evaluations scored lower (`2` and `3`) alongside the fact that there are no `5` evaluation scores. For the male professors, however, there are less of the lower scores and more of the `4` and `5` scores. This seems to show that overall, male professors have better-scored course evaluations.

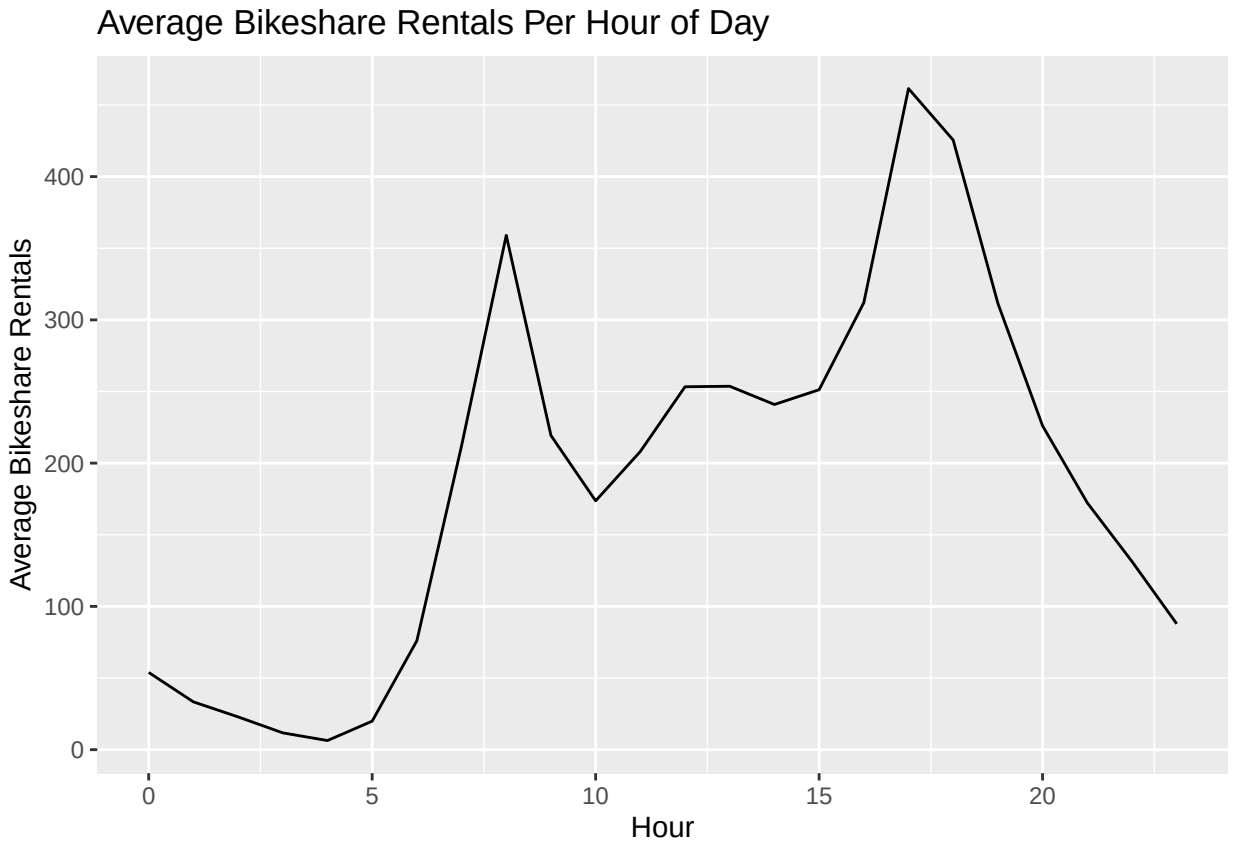
1 - Part D



Looking at the plot itself, the points are scatter everywhere, with no clear pattern to which they are following. Let's test this out with the correlation. The correlation coefficient for this relationship is 0.1890391, which is pretty low (considered a Very Weak Positive correlation). From this, we can understand there is no apparent relationship between how attractive a professor is and the score they receive on the course evaluations.

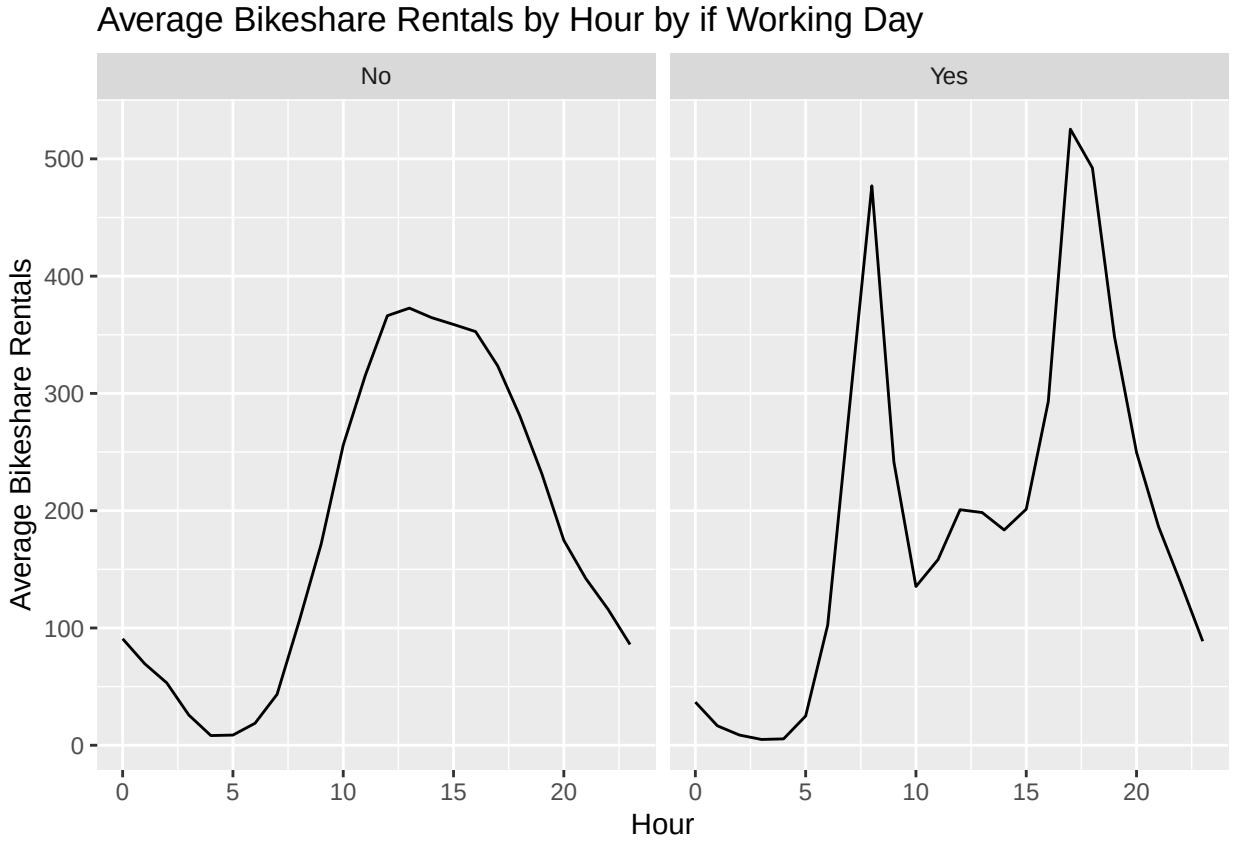
Problem 2 - Bike Sharing

2 - Plot A



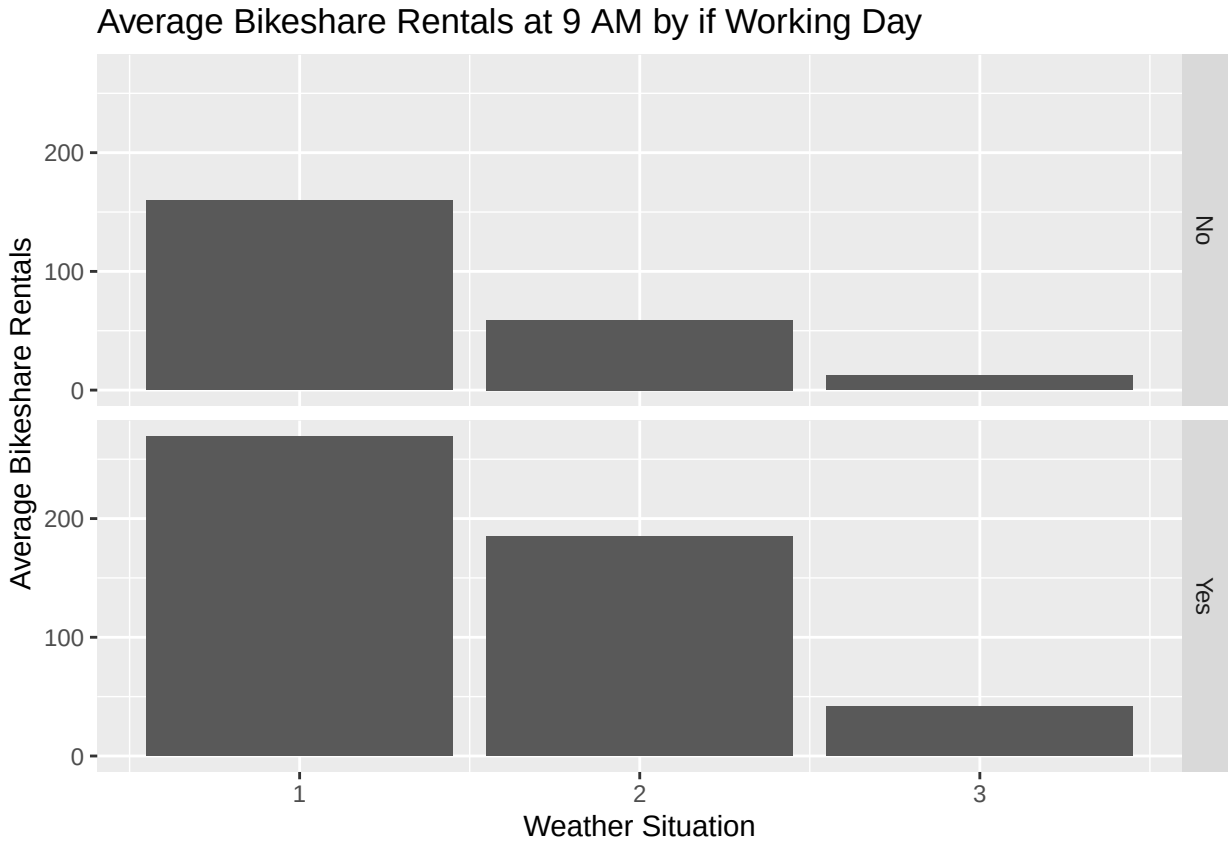
There is are two very visible peaks: one at 08 (8 AM) and another at 17 (5 PM). This is the average workday for most people, a 9 to 5. (8 is the peak as that is travel time to work). The plateau near 12 and 13 is lunchtime. The peak in the evening is higher up compared to the morning, showing that people prefer biking in the evening rather than the morning on their way to work. Since this data was taken from Washington D.C. it makes sense as to why the bikeshare rentals are very high in number (anything to avoid the traffic!). The time when the least amount of bikes are rented out is at 4 AM. In all, we can learn that high bike traffic ocurrs during regular 8 AM and 5 PM times, and more people bike in the evening than in the morning.

2 - Plot B



The graph for when the bikes rented on working days is a lot more spiky then the graph for the bikes rented on non-working days. For the non-working days, the peak of the graph only hits near midday while for non-working days, the same two peaks as before (8 AM and 5 PM) with the 12 PM plateau are seen. The rise in morning usage grows a lot faster on working days compared to non-working days. At the end of the day, however, they both return to around 100 bikes. Another important detail to note is that the volume of bikes rented on non-working days during midnight (12 AM to 4 AM) are much higher than on working days. From this, we can learn that people stay up later during non-working days than working days and more people bike during midday on non-working days compared to working days.

2 - Plot C

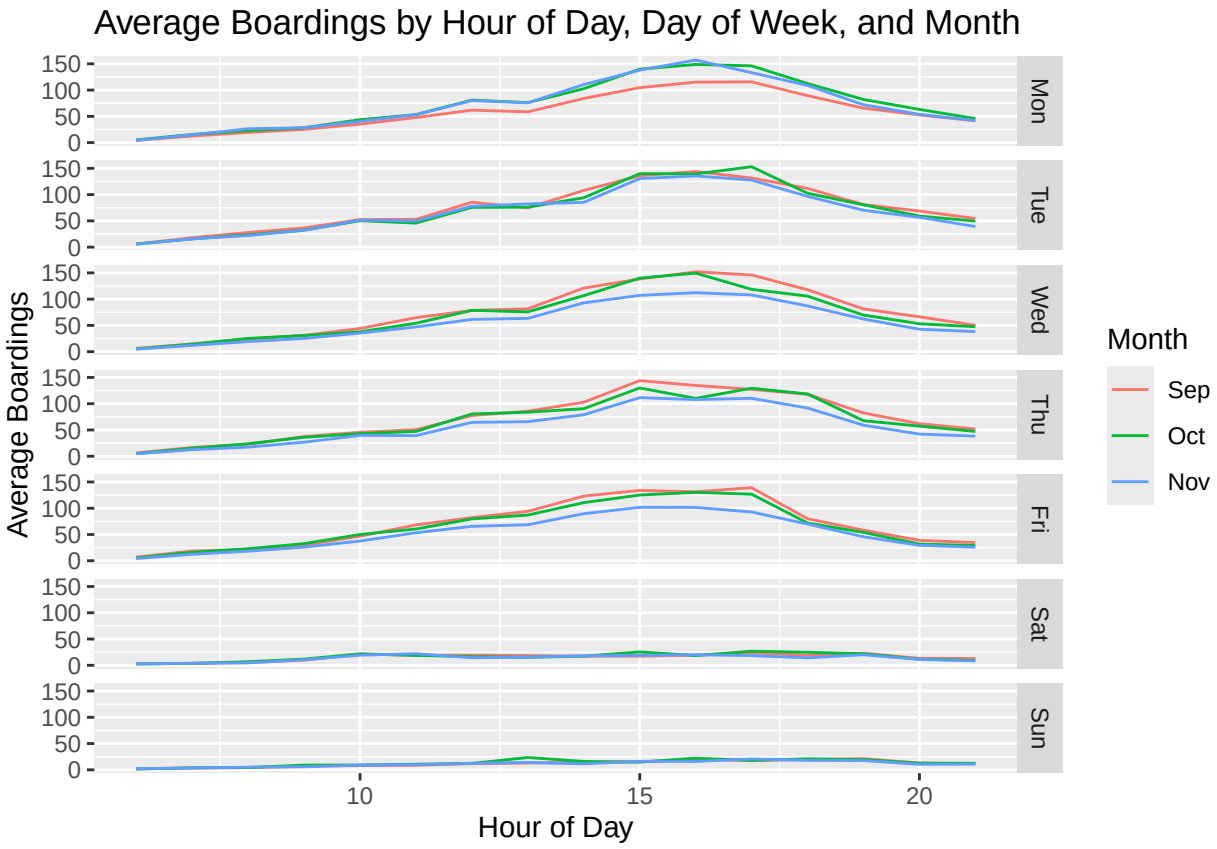


The Weather Situation scale is as follows:
– 1: Clear, Few clouds, Partly cloudy, Partly cloudy
– 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
– 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
– 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

Only the first three weather situations were observed at 9 AM, with `1` being the most observed and `3` being the least observed (which makes sense as clouds are seen in the morning sky more often than snow and rain are). If we look at the data from the previous graph, we can see that the volume of bikes rented out at 9 AM for non-working days is lesser compared to working days. Because this volume during working days is larger, there are more observances of each weather situation. In summary, the bike traffic at 9 AM on working days is larger than on non-working days.

Problem 3 - Capital Metro UT Ridership

3 - Plot 1



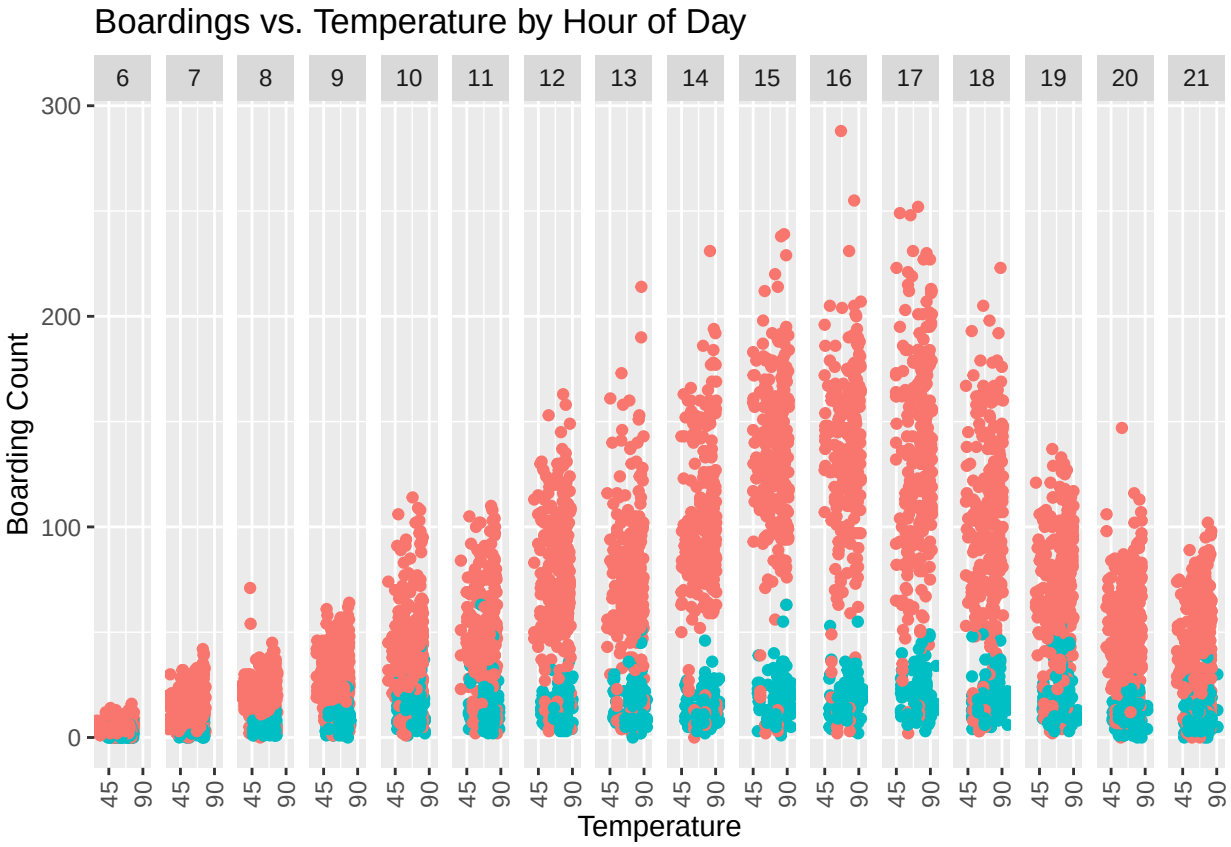
Looking at the graphs, it's clear that the average boardings are greater on the weekdays than the weekends. The most common hours of the day when students ride the CapMetro buses are from 15 (3 PM) to 17 (5 PM). This is a reasonable expectation, since classes end during this time period.

Does the hour of peak boardings change from day to day, or is it broadly similar across days?
As examined before, the weekends and the weekdays are where the main difference lies in boardings. Across the weekdays, the peak boardings happen at relatively the same time. On the weekends, there is no clear picture, as on Saturday is looks like peak boarding time is 15 (3 PM) and on Sunday it is at 13 (1 PM). When examining this overall, it does seem to be broadly similar across the days.

Why do you think average boardings on Mondays in September look lower, compared to other days and months?
Looking at the news article here (https://thedailytexan.com/2018/09/14/students-face-overcrowding-on-buses-following-capmetro-remap/), there seems to have been rerouting of the bus routes in September of 2018, and overcrowding became an issue. This would have caused the students to change their mode of transport. Near the end of the article, Mariette Hummel (CapMetro communications specialist) states that they were to implement changes in the routes (and were already making short-term adjustments to routes). An article from early November 2018 (https://thedailytexan.com/2018/11/05/capmetro-adds-buses-to-route-670-to-help-with-overcrowding/) states that buses were added to route 670 to mitigate the overcrowding issue, which could be why students returned to the buses in October.

Similarly, why do you think average boardings on Weds/Thurs/Fri in November look lower?
There are several reasons for this. Thanksgiving Break is a massive reason (a week of low bus usage, drops a weekly count for each day). Seniors who don't have many classes before the break may just leave early for the long weekend before the break as well, further dropping bus usage.

3 - Plot 2



For easier visualization, I have removed the legend. The orange (or salmon) dots are for `weekday` and the teal dots are for `weekend`.

Observable is a general trend where as the day passes (and this is true for both the weekdays and weekends), the peak time for bus usage is the 15 (3 PM) and 16 (4 PM) times, similar to what we observed in the previous plot. In each of the graphs, there seems to be a slight upward trend as we make it through the temperatures, showing boarding has a positive correlation with temperature (with higher temperatures come more students boarding buses).

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?
Yes. When we look at each hour, there is the general upward trend of points (meaning with higher temperatures per hour there is a larger number of students boarding the bus).

Problem 4 - Wrangling the Billboard Top 100

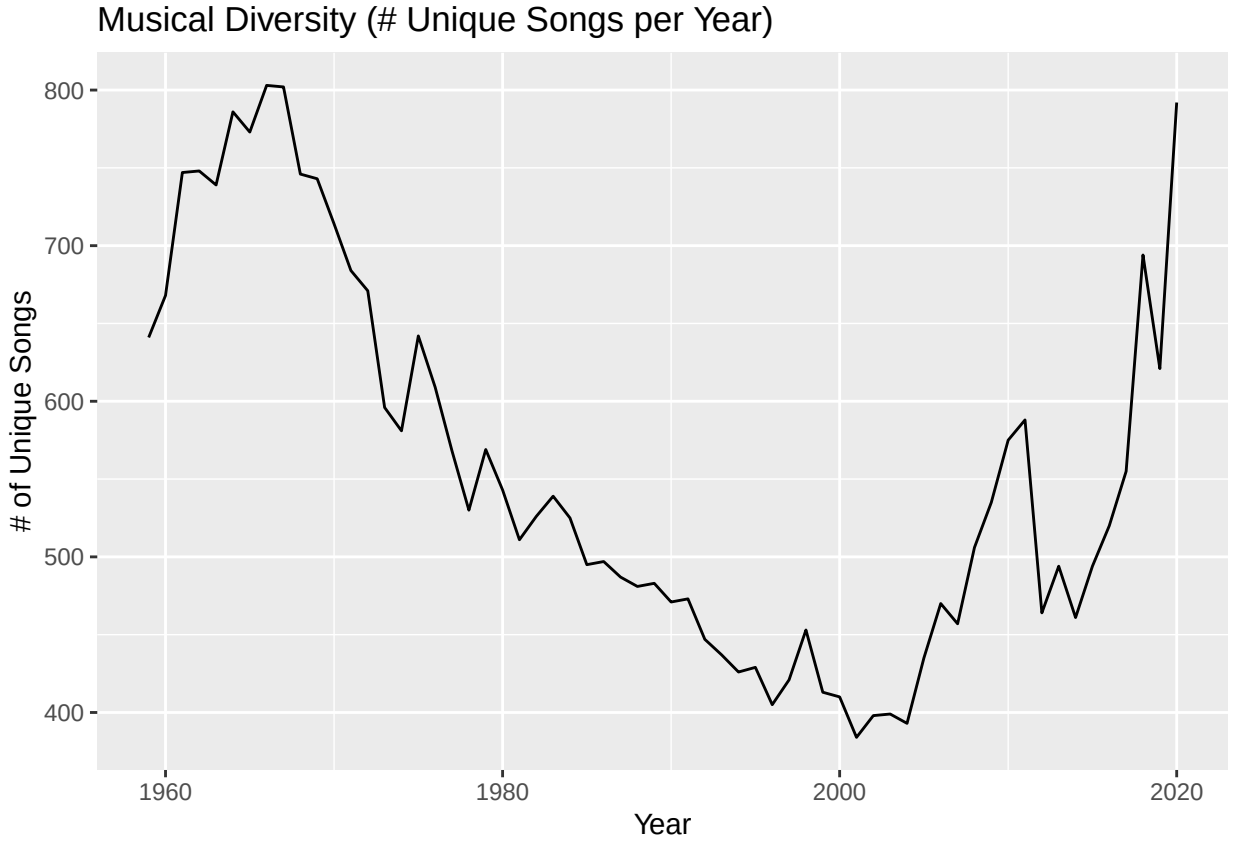
4 - Part A

`Top 10 Most Popular Songs on Billboard Top 100 (1958-2021)`

performer	song	count
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79
Jason Mraz	I'm Yours	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
Lifehouse	You And Me	62

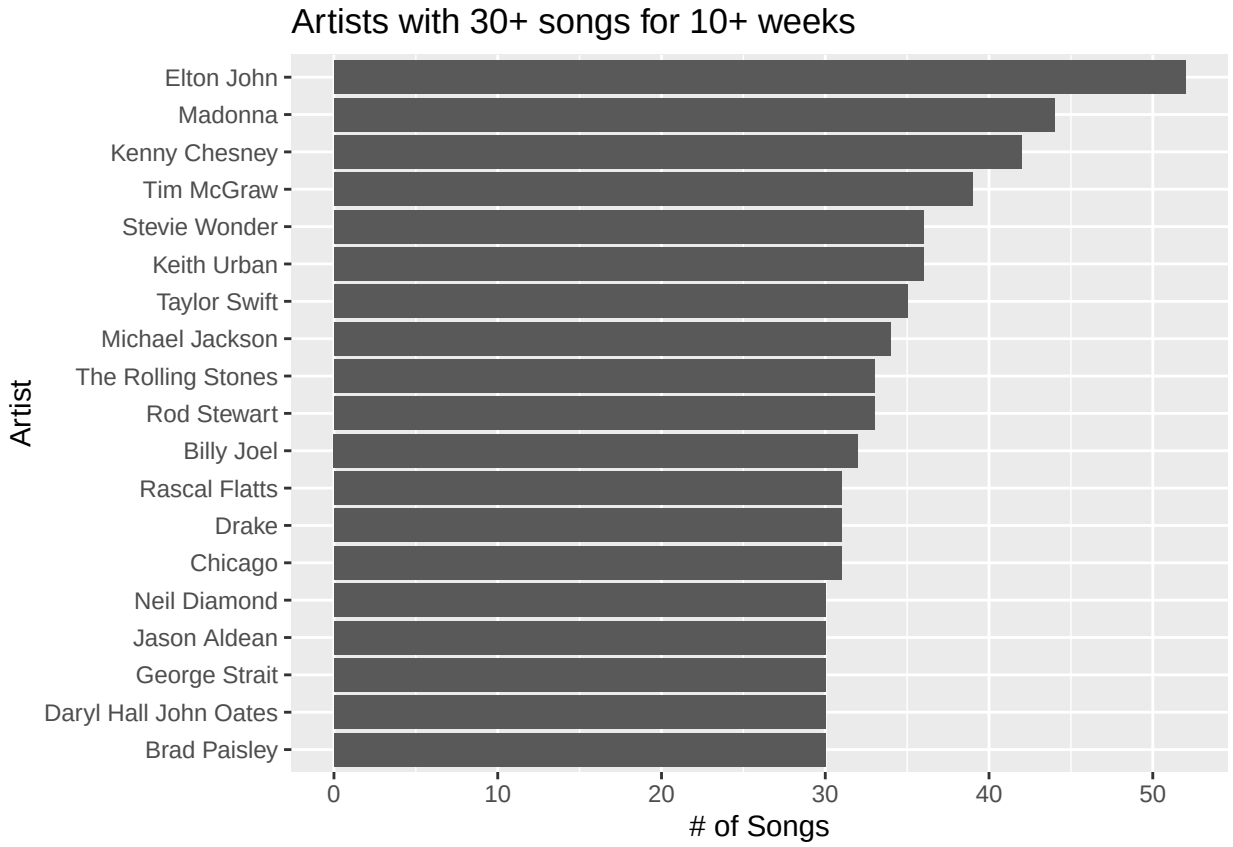
Here, we can see the top 10 songs from the years 1958 to 2021 that lasted in the Billboard Top 100 the longest, with the longest-standing song "Radioactive" by Imagine Dragons (having 87 weeks under its belt). In second place comes "Sail" from AWOLNATION with 79 weeks under its belt. The 10th longest lasting song on the Billboard Top 100 is "You And Me" by Lifehouse, having lasted 62 weeks.

4 - Part B



Immediately we see a decline in the number of unique songs that entered the Billboard Top 100 per year from the late 1960s to the mid 2000s. From the mid 2000s, however, we started seeing the revival in unique songs, and in 2020 we are approaching the same number of unique songs we had in the late 1960s. Musical diversity is now making a comeback.

4 - Part C



Here, as the title of the plot explains are the 19 artists, since 1958, who have had at least 30 songs on the Billboard Top 100 for more than 10 weeks (essentially 30 or more "ten-week hits"). The plot above shows the performers (artists) and the number of songs they have had as "ten-week hits". Elton John comes out on top, with 52 songs. Neil Diamond, alongside Jason Aldean, George Strait, Daryl Hall John Oates, and Brad Paisley, are all tied in last place with 30 songs exactly.