Maximizing Enjoyment: Reducing Wait Times at Walt Disney World

GROUP 4:
JACOMO CORRIERI,
Kim Minyeong,
Lee Jimin

OUTLINE OF THE PROJECT

01
Introduction

02
Preliminary Findings

03
EDA Part 1

04
EDA Part 2

05
Challenges

06
Conclusions

Introduction

Q: Which day should I visit Disneyland to be most likely to have good weather and short waiting times?

What's the data?

- Disneyland Visitors Data (2018-2022)
  - Various information on **rides** given over 15 minute intervals
  - As well as **weather** conditions over 1 hour intervals
- Available via Kaggle in .csv format
- Primary datasets include:
  - waiting_time.csv **3.5M** Rows (Observations) and **14** Columns (Variables)
  - weather_data.csv **207k** Rows (Observations) and **28** Columns (Variables)
  - Very large datasets that would need to be trimmed (3.5M * (14+28) ~= 147M!)

※ There is a abbreviation explanation on the last slide

Preliminary Findings

Dataset Features

- Key Columns - waiting_times.csv
  - **Date, Time**, Ride Name, Ride Information
    - wait time max: Max waiting time during the considered period (in minutes)
    - Capacity: Capacity of the attraction
    - open time: Open time of the attraction (in minutes)
- Key Columns - weather_data.csv
  - **Date, Time**, Temperature, Humidity, Wind, Rain, Weather Description

Data Cleaning and Combining

1. Remove unique variable columns, highly-related columns
2. Combining by Date and Time (1 hour intervals)
   - Different time spans (Waiting_times: 2018-2022, Weather: 1970-2022)
3. Categorize time into ‘Morning’, ‘Afternoon’ and ‘Evening’

EDA and Initial Questions

1 . Combined Dataset

We now have a combined dataset featuring substantial ride and weather data.

- Combined Dataset - **136,536** Rows x **24** Columns

2. Main Question

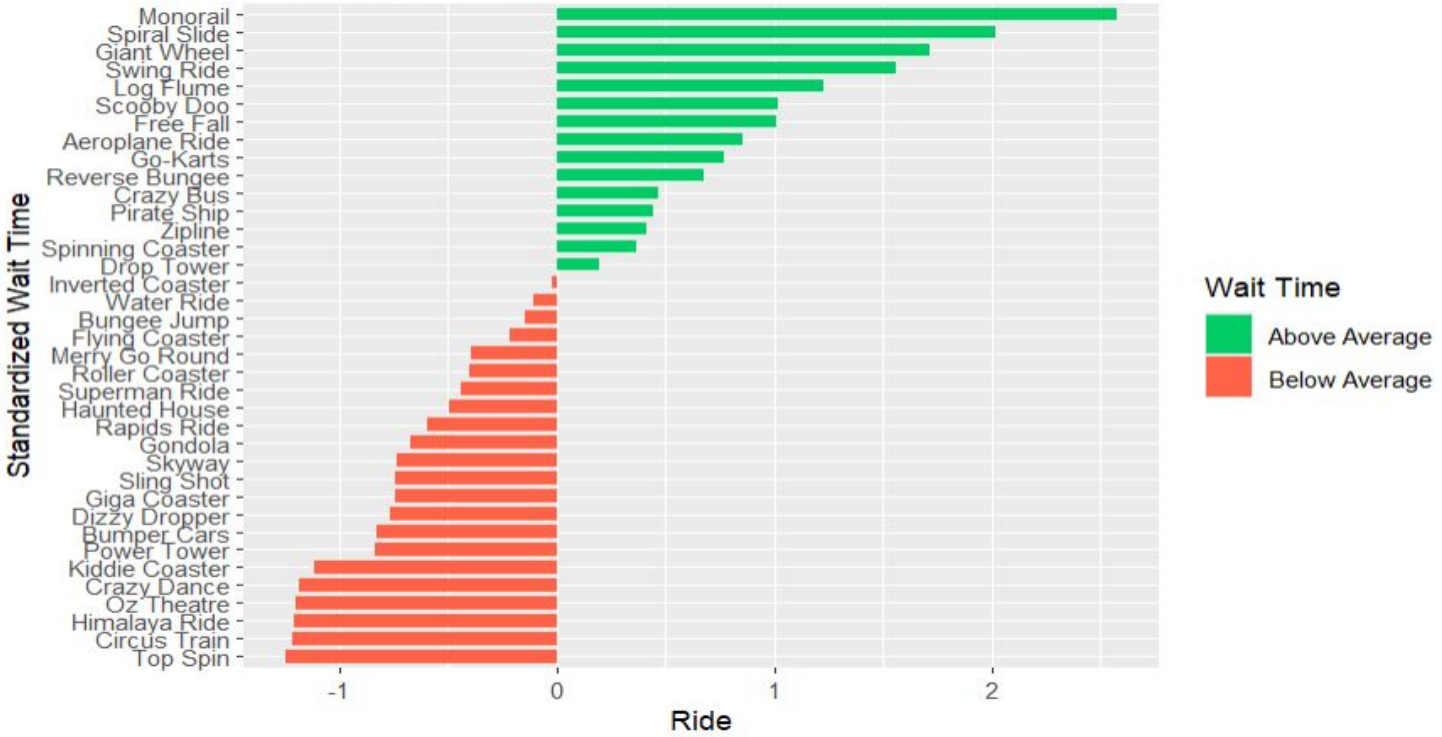
Suppose that Jimin and Minyeong want to visit Jacomo in Florida to go to Disney World. When would be the best time to come? During which conditions? Which rides should they ride, and when would be the best time of day?

3. Initial Questions

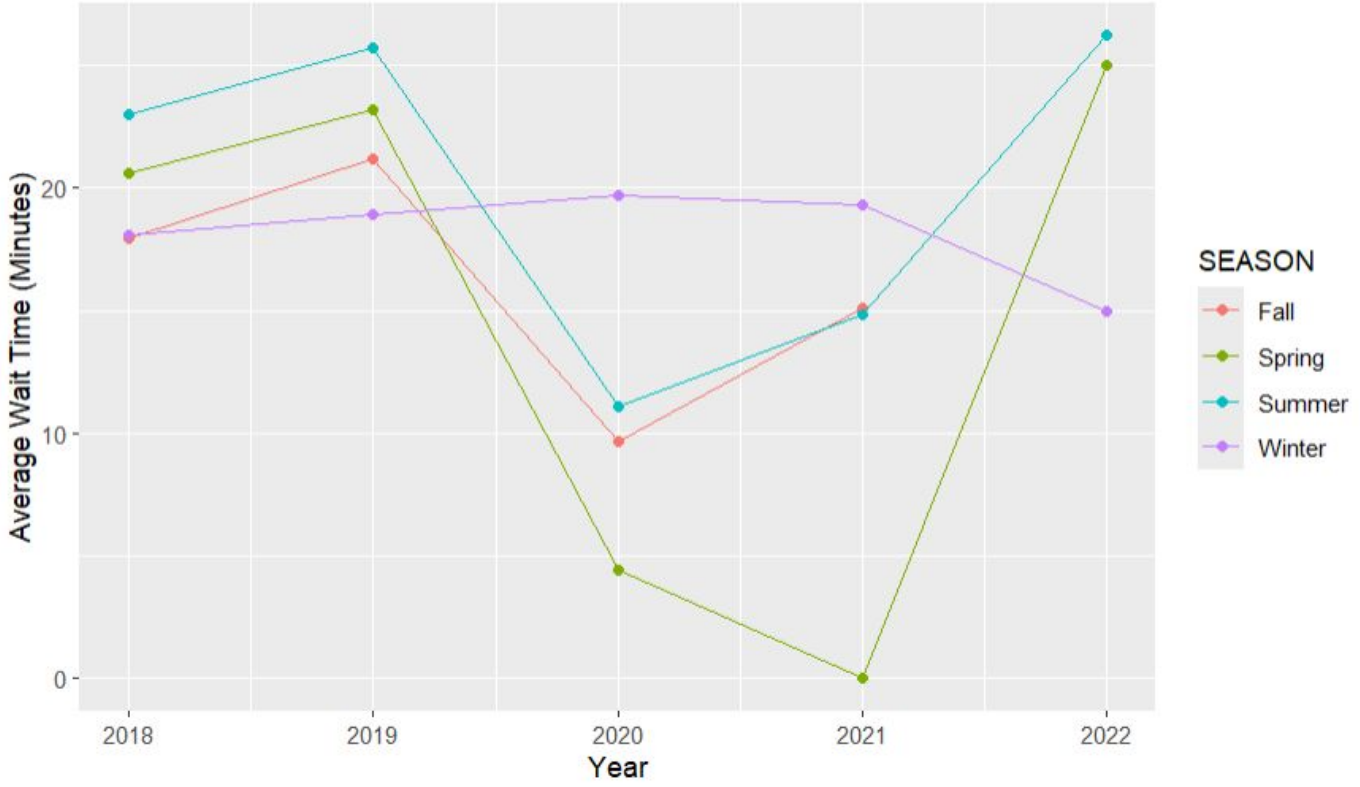
What variables have a visible linear effect on the average waiting times?

During which seasons, months, and days is the park busiest?

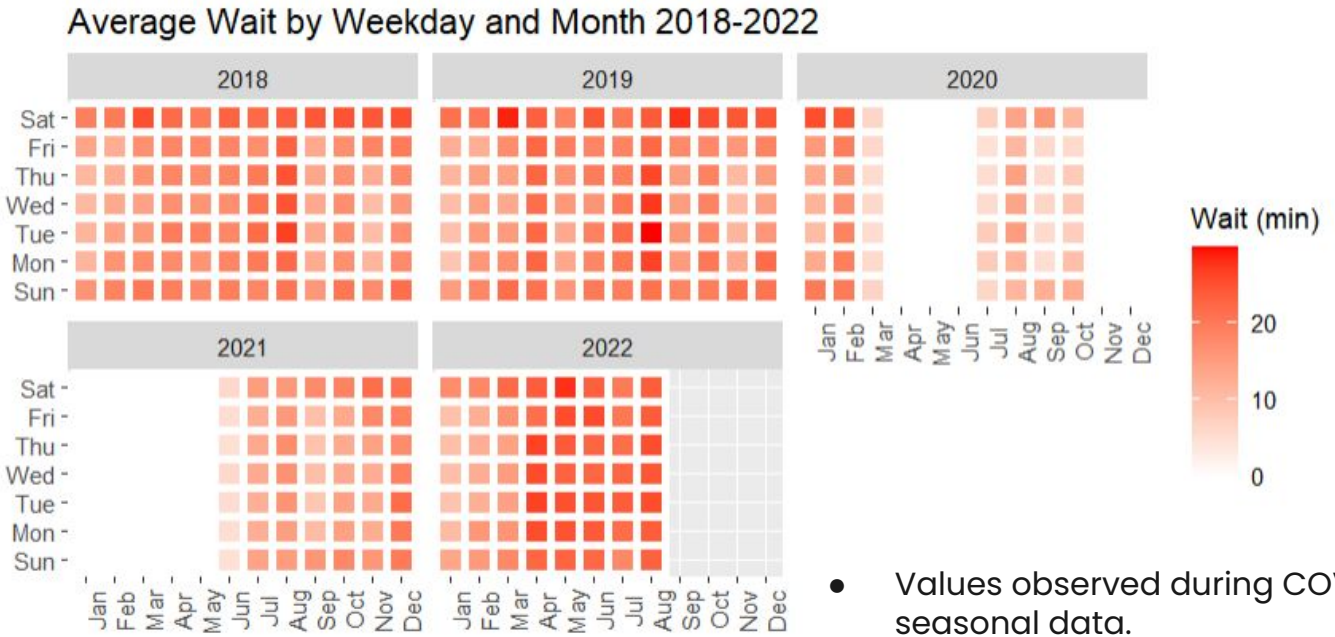
Waiting_Times (by Ride)



Waiting_Times (by Year)

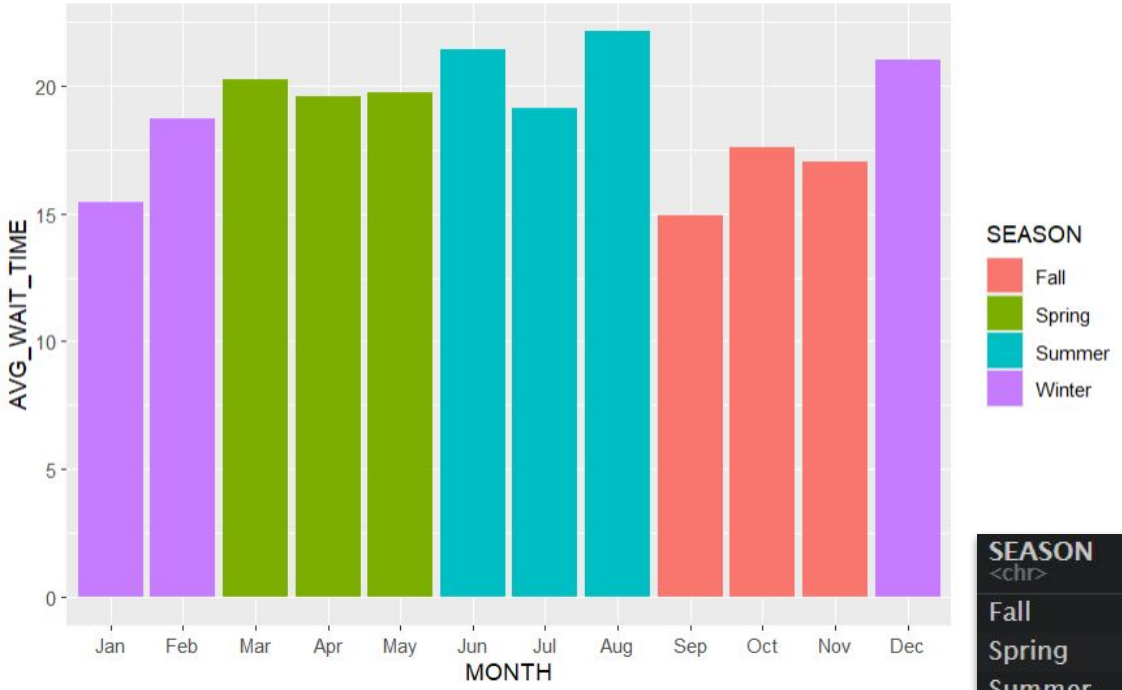


COVID-19's Effect on Data



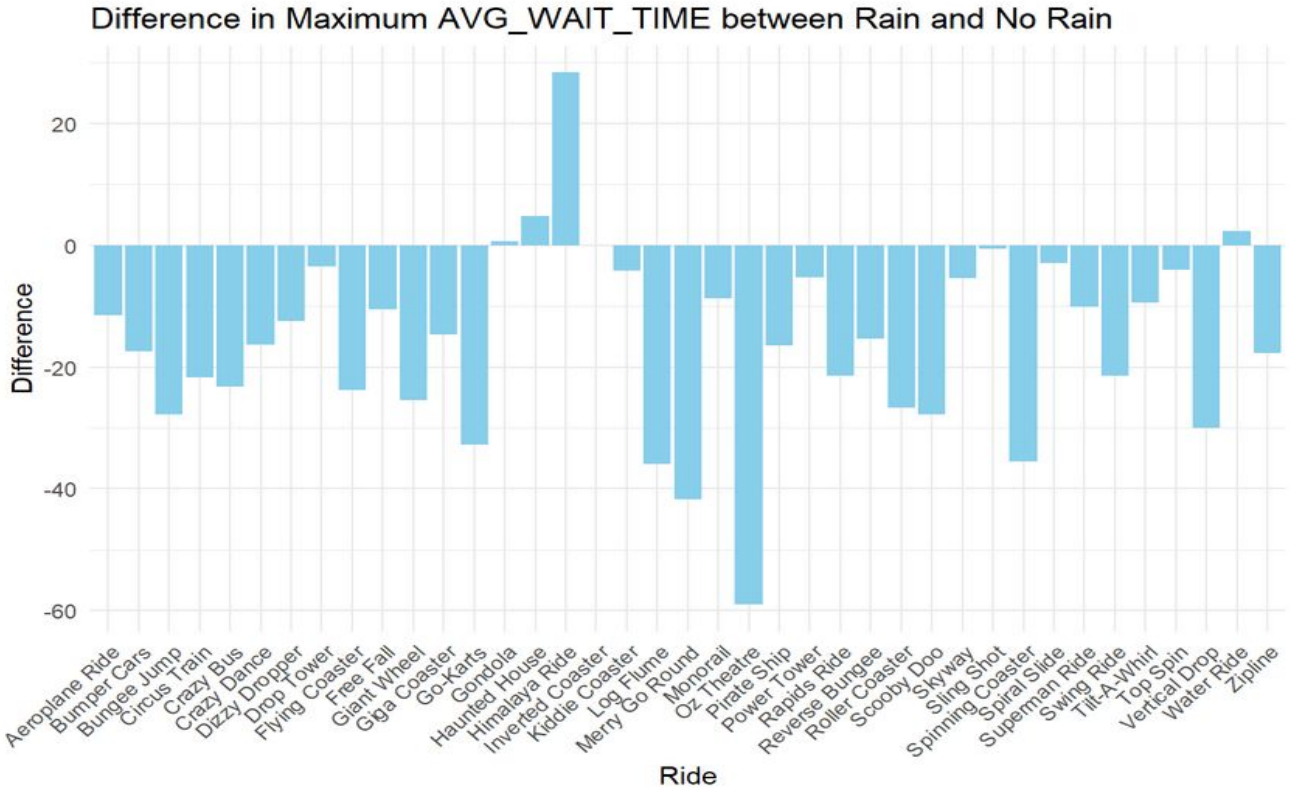
- Values observed during COVID (0s) could skew seasonal data.
- Removed observations holding zero as the value for every numeric column.

Which Season would be best?

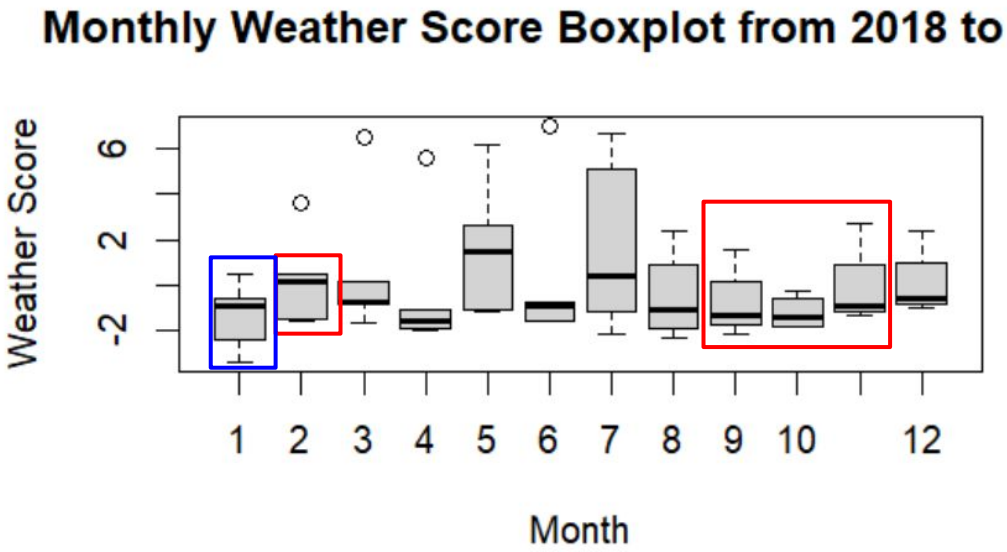


SEASON <chr>	Average_Wait <dbl>
Fall	16.47582
Spring	19.86223
Summer	20.83279
Winter	18.19492

The Effect of Rain on Ride Waits



Determining the Ideal Month



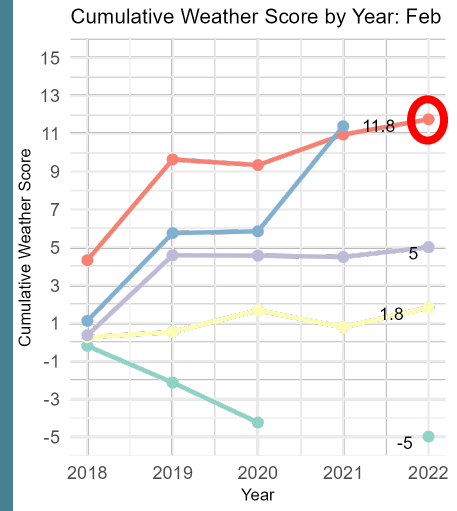
The scores of each descriptions are depending on the subject. If the weather criteria is different according to the subject, the score could be changed.

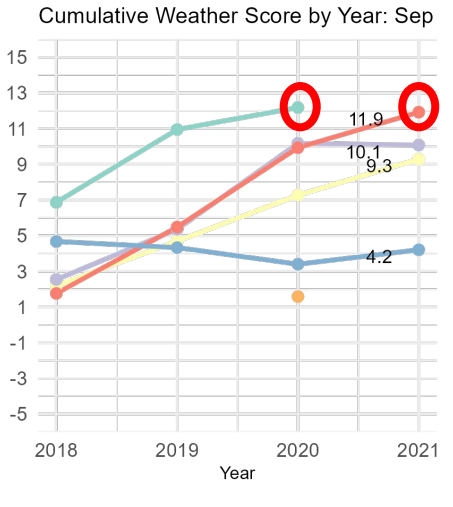
By **weather_description** of weather_data.csv

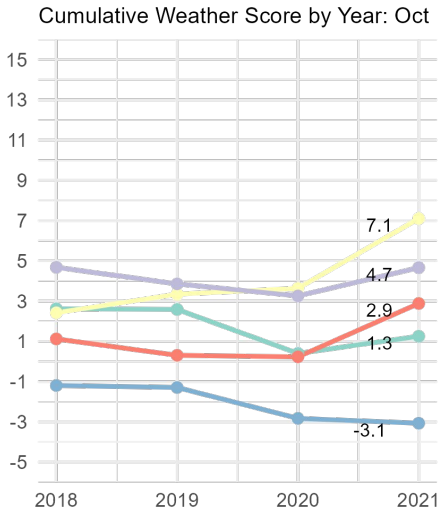
-
- **sky is clear** = 7
- **few clouds** = 5
- **scattered clouds** = 3
- **broken clouds** = 1

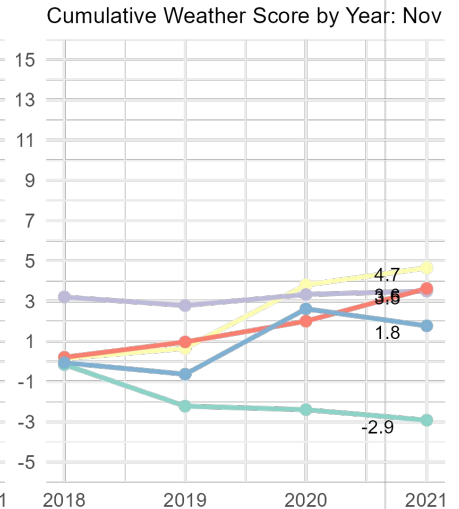
- **overcast clouds** = -1
- **light rain** = -2
- **moderate rain** = -4
- **light snow** = -3
- **snow** = -5
- **heavy intensity rain** = -10

Finding the Best week - Wednesdays, Nov

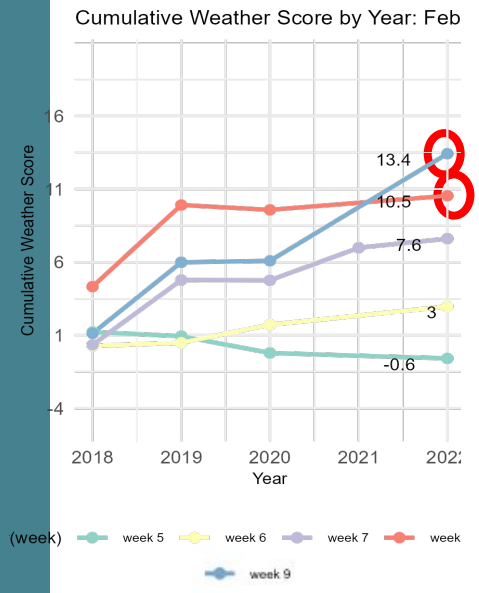


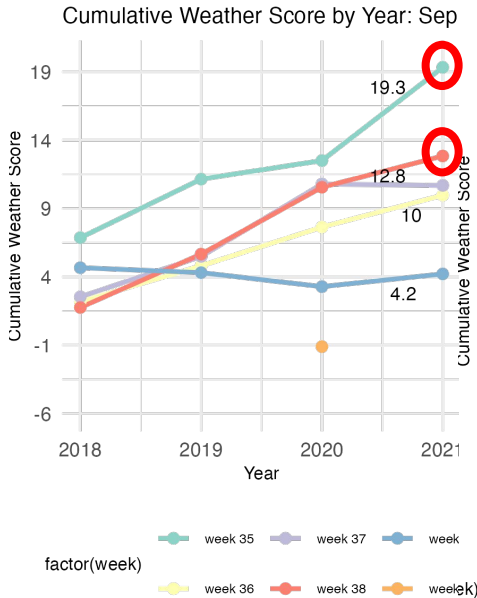


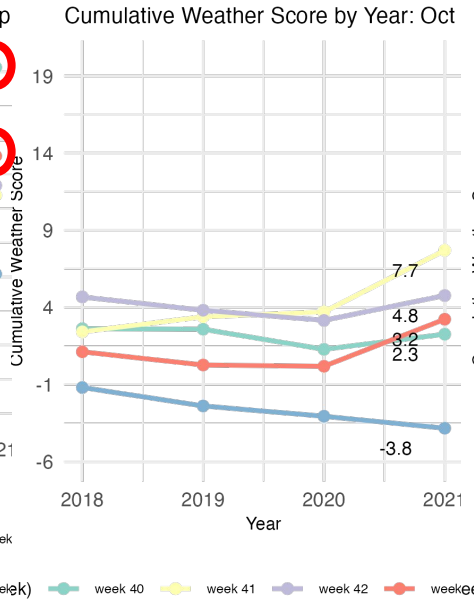


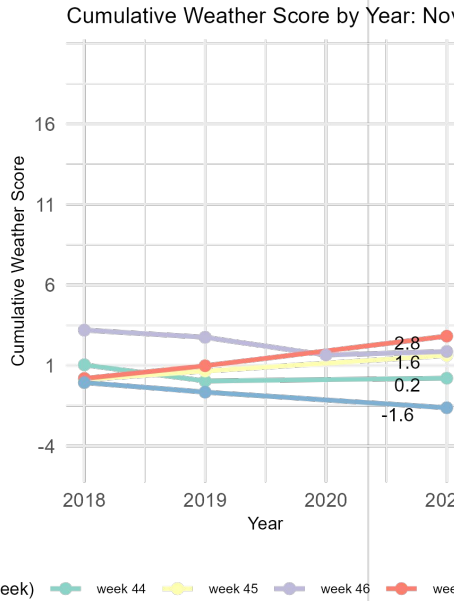


Finding the Best Month and Week

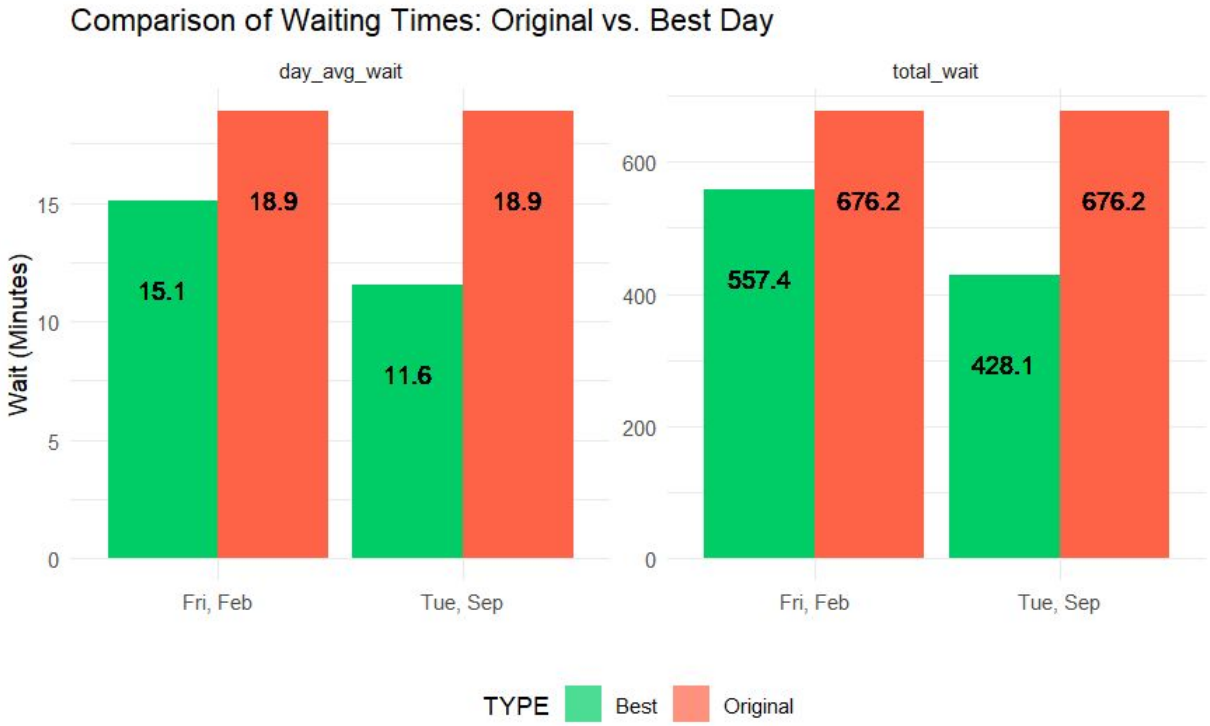




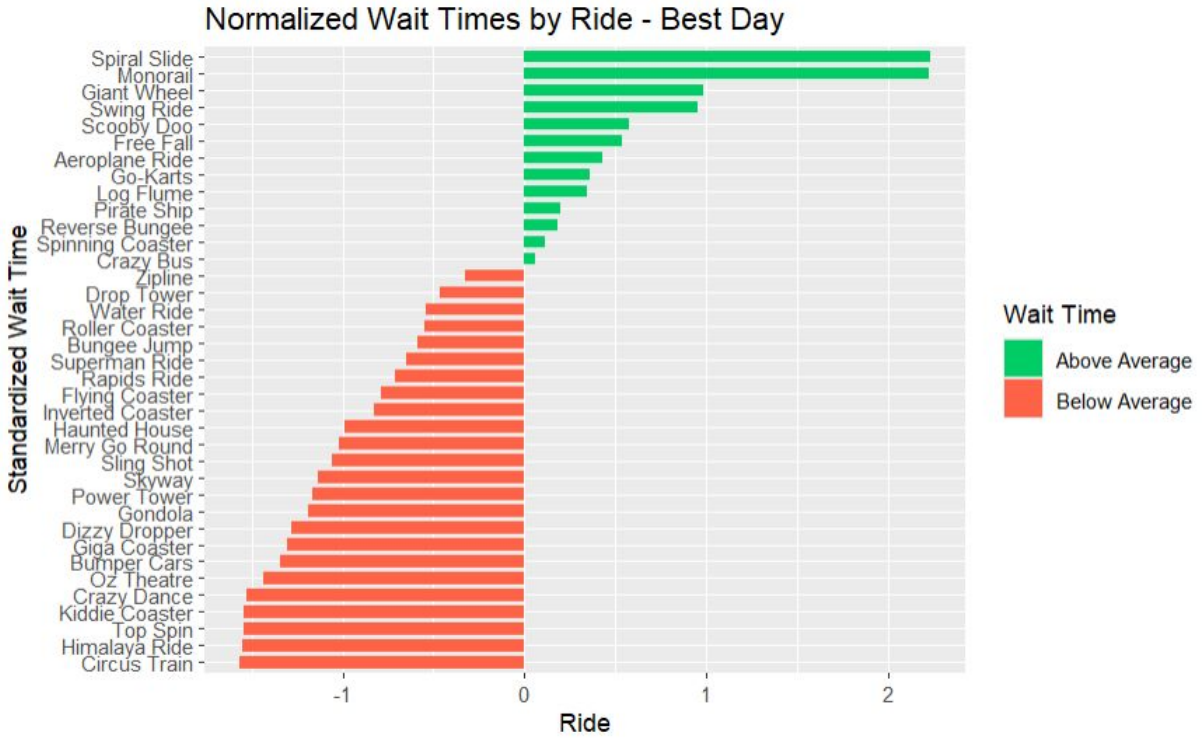




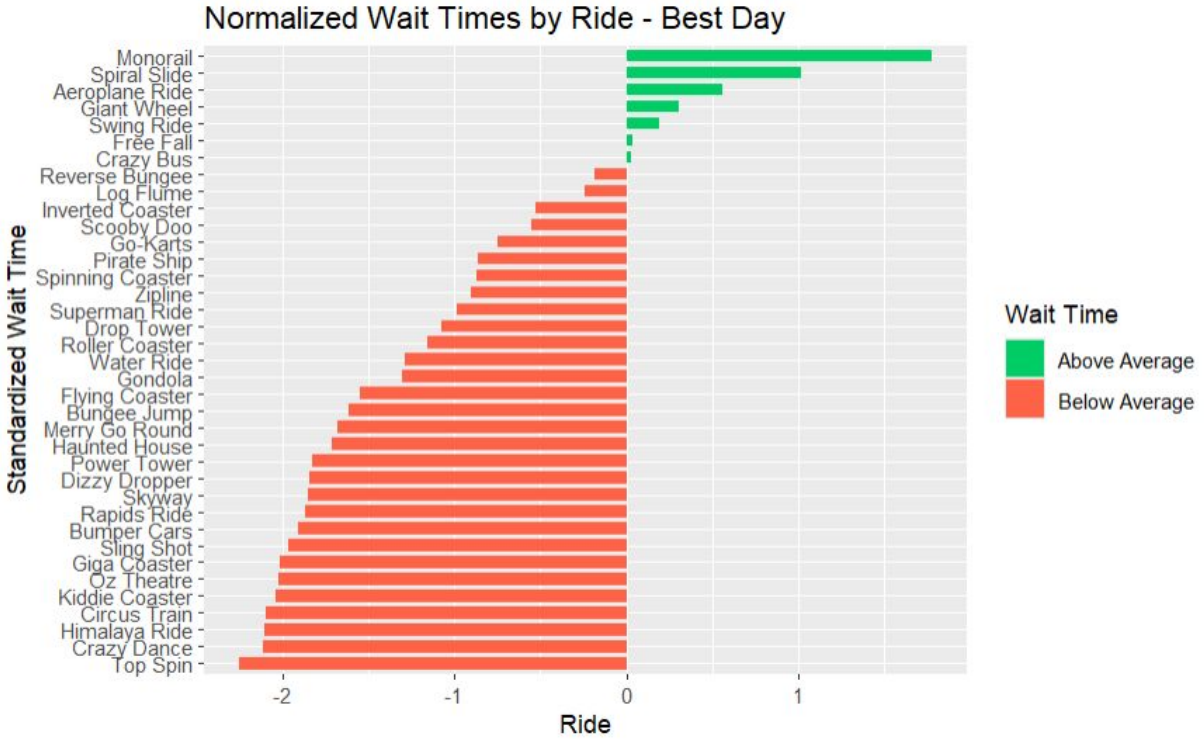
Finding the Best Day



Finding the Best Day – Fridays, Feb



Finding the Best Day – Tuesdays, Sep



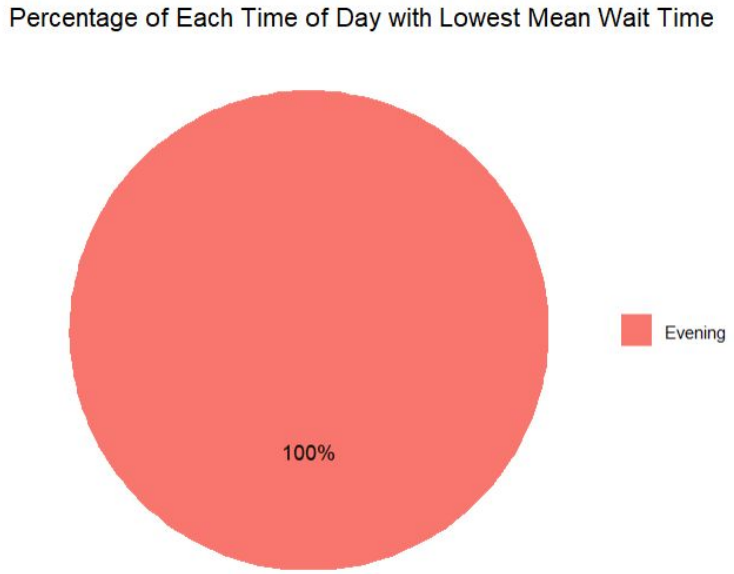
Top Five Rides - Best Time of Day

Friday, February

	RIDE	MEAN_WAIT	WDAY	MONTH
	<chr>	<dbl>	<ord>	<ord>
1	Spiral Slide	43.4	Fri	Feb
2	Monorail	43.2	Fri	Feb
3	Giant Wheel	30.0	Fri	Feb
4	Swing Ride	29.8	Fri	Feb
5	Scooby Doo	25.8	Fri	Feb

Tuesday, September

	RIDE	MEAN_WAIT	WDAY	MONTH
	<chr>	<dbl>	<ord>	<ord>
1	Monorail	33.1	Tue	Sep
2	Spiral Slide	27.4	Tue	Sep
3	Aeroplane Ride	23.9	Tue	Sep
4	Giant Wheel	21.9	Tue	Sep
5	Swing Ride	21.1	Tue	Sep

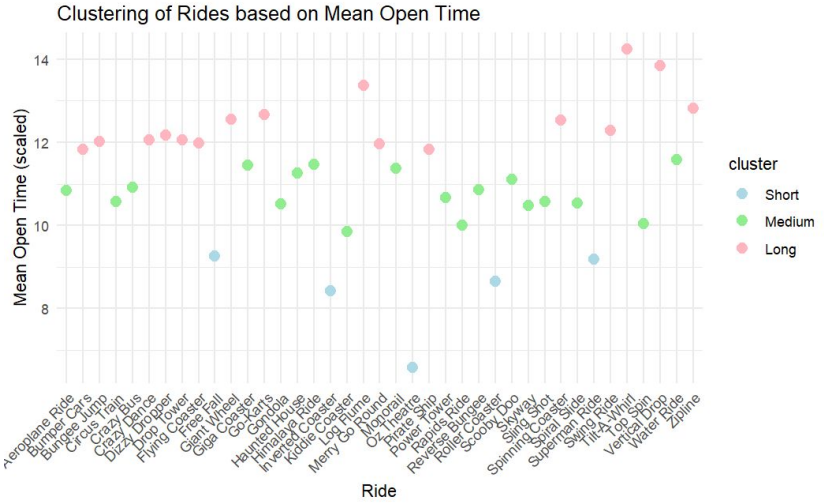


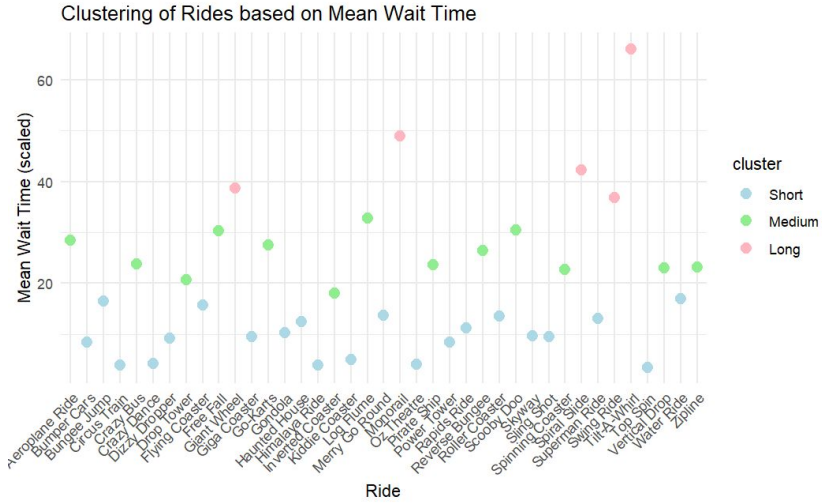
Correlation between variables

- OPEN_TIME : ride's operating time

```
Call:
lm(formula = OPEN_TIME ~ AVG_WAIT_TIME, data = open_main_df)

Residuals:
    Min      1Q  Median      3Q     Max
-23.019  -2.993   1.167   3.633   6.415

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.5846720  0.0168754   508.7   <2e-16 ***
AVG_WAIT_TIME 0.1312178  0.0006148   213.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.515 on 136534 degrees of freedom
Multiple R-squared:  0.2502,    Adjusted R-squared:  0.2502
F-statistic: 4.555e+04 on 1 and 136534 DF,  p-value: < 2.2e-16
```

Can we see the actual correlation?





- Grouping by RIDE, calculated the average of Open time and Wait time for each group
- Standardized the average Open time and Wait time of each ride
- Creating **3** clusters with **K-means Clustering**

-> Each cluster : Rides with similar mean Open time / Wait time

Can we see the actual correlation?

```r
cluster_1_wait <- clustered_rides_wait$rides_in_cluster[clustered_rides_wait$cluster == 1]
cluster_3_open <- clustered_rides_open$rides_in_cluster[clustered_rides_open$cluster == 3]
matched_rides1 <- intersect(cluster_1_wait, cluster_3_open)
matched_rides1
```

```
character(0)
```

Cluster_1_wait: Shortest waiting time cluster / Cluster_3_open: Longest open time cluster

- **Hypothesis**

  Since there is a correlation of 0.50, rides belonging to both clusters with shortest waiting times and clusters with longest open times would be derived.

- **Result**
  There were no matching rides between two clusters. Hard to see the actual correlation.
  (Both longest waiting time cluster & shortest open time cluster
  longest open time cluster & shortest waiting time cluster showed no matching rides)

Challenges

1 . Trimming the data

- Excluding data during COVID
- Removing columns that are overlapping to the other
- Categorizing time into morning, afternoon, evening

2. Reaching to the result

- Narrowing down to from big to small

Conclusions

- Rain tends to result in decreased waiting times
- Open time has the strongest correlation with waiting times, though it is still a fairly weak relationship.
- Jimin and Minyeong would prefer to visit WDW on a **Tuesday in September** on either week 35 or 38, but may have to come during **February on Friday** on week 9(Winter Break, second best option).

THANKS!

Do you have any questions?

※ The descriptions are referred to ‘glossary’ file of the dataset

Variable Explanation: waiting_times.csv

WORK_DATE	Date of the event
DEB_TIME (FIN_TIME)	Start(End) time of the event
DEB_TIME_HOUR	Start hour of the event
ENTITY_DESCRIPTION_SHORT	Name of the entity (park, attraction)
WAIT_TIME_MAX	Max waiting time during the considered period (in minutes)
NB_UNITS	Number of units in the attraction
GUEST_CARRIED	Number of guest carried during the period
CAPACITY	Capacity of the attraction
ADJUST_CAPACITY	Adjusted capacity of the attraction to the time slot considered
OPEN_TIME (UP_TIME, DOWNTIME)	Open(up, down) time of the attraction (in minutes)
NB_MAX_UNIT	Number max of units that the attraction can take

Variable Explanation: weather_data.csv

dt	Unix timestamp (seconds since 1970-01-01 00:00:00 UTC)
dt_iso	Date and time in ISO 8601 format (YYYY-MM-DD HH:MM)
timezone	Timezone offset from UTC in seconds
temp	Temperature at the given time (in degrees Celsius)
dew_point	Dew point temperature (in degrees Celsius)
feels_like	Perceived temperature (in degrees Celsius)
clouds_all	Cloudiness percentage
weather_id	Weather condition ID (corresponding to a specific weather condition)
weather_main	Main weather condition (e.g., Clear, Clouds, Rain)
weather_description	Detailed description of the weather condition
weather_icon	Weather icon ID (used for visual representation of the weather)

Weather Data

Preparing the Data

- Eliminating redundant columns
- Formatting the date and time
  - There is an entry for each hour of the day, so 24 entries per day
  - We can simplify this by time of day (morning, afternoon, evening, night) [9-13), [13-18), [18-23), [23-9)
- Numeric values were averaged across the three time periods
- Trimming the date range to match with Waiting Times
  - Waiting Times (2018-2022)
  - Weather Data (1970-2022)
- Creating categories based on the given weather description
- Trimmed down to 6,764 Observations

Waiting_Times

Preparing the Data

- There are 39 unique rides
- Two rides were added later (June 2022), so a majority of the dataset uses 37 rides
- ~2072 entries per day, so (2072 entries) * (1,691 rows) = 3.5M obs.
- We can condense the data by time of day (morning, afternoon, evening, night) [9-13), [13-18), [18-23), [23-9)
    - We have specific time data (h/m/s) for each ride's observation
- No data was recorded during the "night" period, so this category was filtered out
- Each time period spans 4-5 hours (for even distribution)
- All ride information was averaged for each time of day
- 3 entries * 39 rides * 365 days * 5 years (18,19,20,21,22) = ~213,525 obs.

Waiting_Times (by Month)