DSC 530 – Data Exploration and Analysis

Steven Miller

Statistical Question

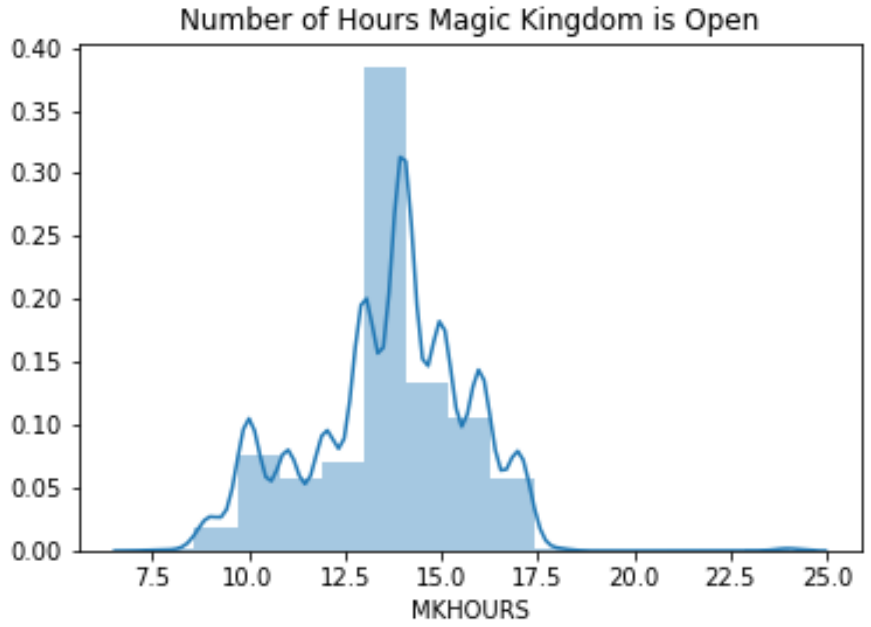
- What factors can we use to predict the wait time of a theme park ride?

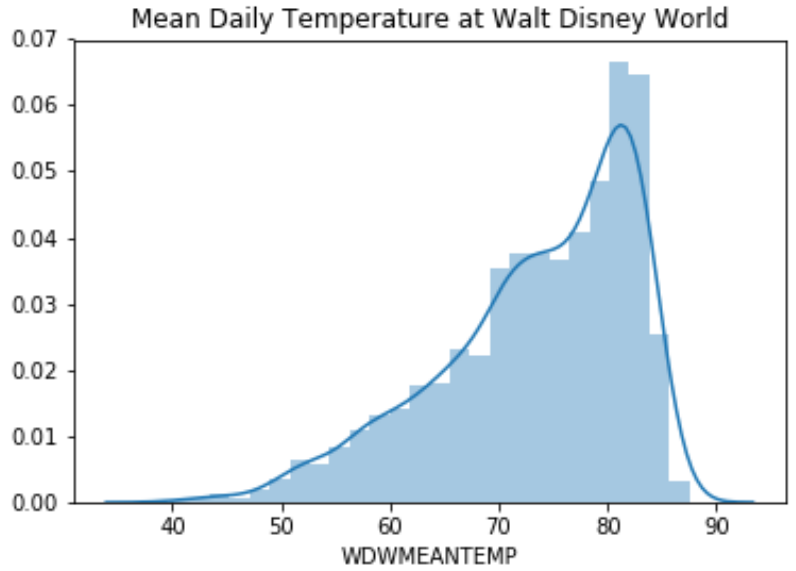
Dataset and Variables Used

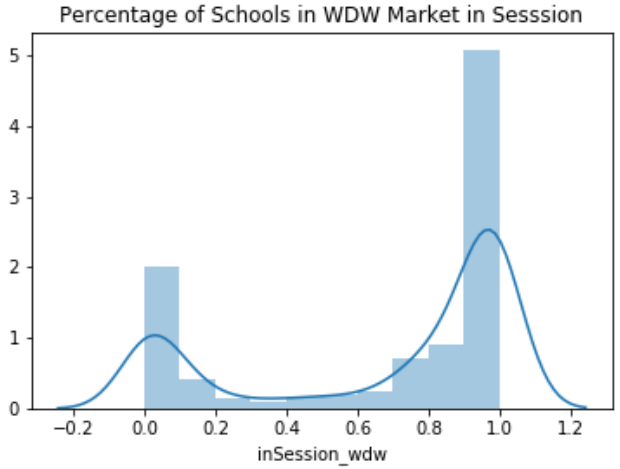
- TouringPlans Splash Mountain data will be explored
  - https://touringplans.com/walt-disney-world/crowd-calendar#DataSets
  - Metadata common to all datasets will also be used.
  - From Splash Mountain data – Date, Datetime and SPOSTMIN
    - SPOSTMIN is the posted wait time in minutes
  - From Metadata:
    - DATE
    - DAYOFWEEK
    - MKHOURS – the number of hours the Magic Kingdom was open on a day
    - HOLIDAYM – a 0-5 scale of holiday seasons, 5 being the highest (Christmas) and 0 being no holiday at all
    - WDWMEANTEMP – average temperature of the day
    - WEATHER_WDWPRECIP – precipitation recorded for the day
    - inSession_wdw – percentage of schools within the Walt Disney World audience that are in session

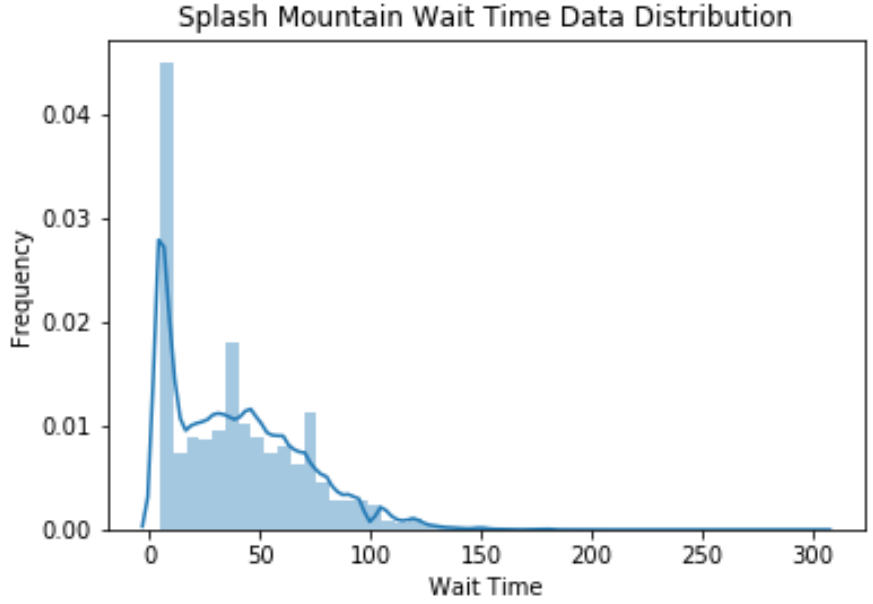
Hypothesis

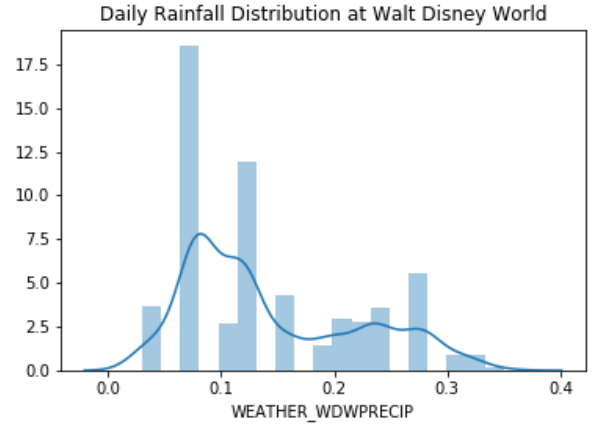
- I expect to find:
  - A positive relationship between weekends and wait times
  - A positive relationship between mean temperature and wait times
  - A positive relationship between holidays and wait times
  - A negative relationship between rainfall and wait times
  - A negative relationship between the percentage of schools in session and wait times
  - No relationship between the number of hours the park is open and wait times.


Number of Hours Magic Kingdom is Open
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
MKHOURS


Mean Daily Temperature at Walt Disney World
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
40
50
60
70
80
90
WDWMEANTEMP


Percentage of Schools in WDW Market in Sesssion
5
4
3
2
1
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
inSession_wdw


Splash Mountain Wait Time Data Distribution
0.04
0.03
0.02
0.01
0.00
Frequency
0
50
100
150
200
250
300
Wait Time


Daily Rainfall Distribution at Walt Disney World
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.0
0.1
0.2
0.3
0.4
WEATHER_WDWPRECIP

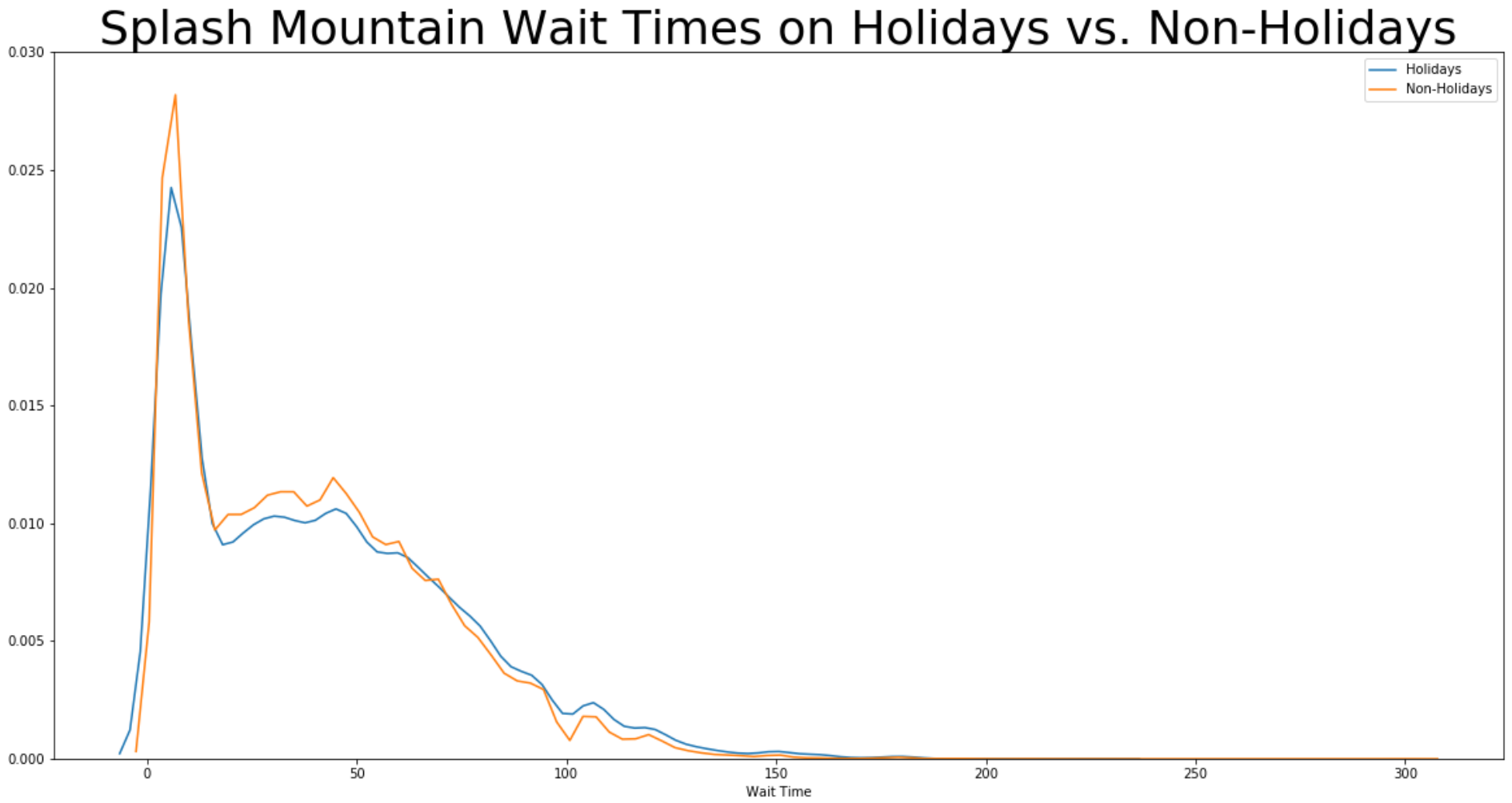
Summaries

Variable	Min	Max	Mean	Median	Std. Dev.	Skew	Kurtosis
SPOSTMIN	5	300	39.45	35	30.09	0.829	0.476
MKHOURS	7.5	24	13.63	14	2.05	-.204	.387
WDWMEANTEMP	39.75	87.49	73.44	75.16	8.93	-.878	.164
WEATHER_WDWPRECIP	0.03	0.35	.144	.12	.078	.719	-.684
inSession_WDW	0	1	.679	.9	.39	-.866	-1.018

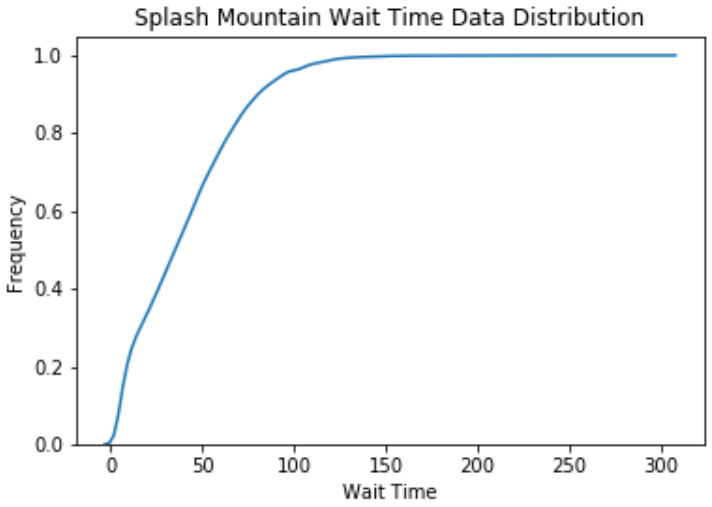
Outliers

- Few outliers are present in these variables, however two stand out.
  - The number of hours the Magic Kingdom has a maximum value of 24, over five standard deviations above the mean. This is a result of an annual event that was previously held where the park was open from 6 AM to 6 AM.
  - The posted wait time goes as high as three hours, despite a mean value of only 39.45. These are likely very extreme and unusual circumstances.

Holidays vs. Non-Holidays

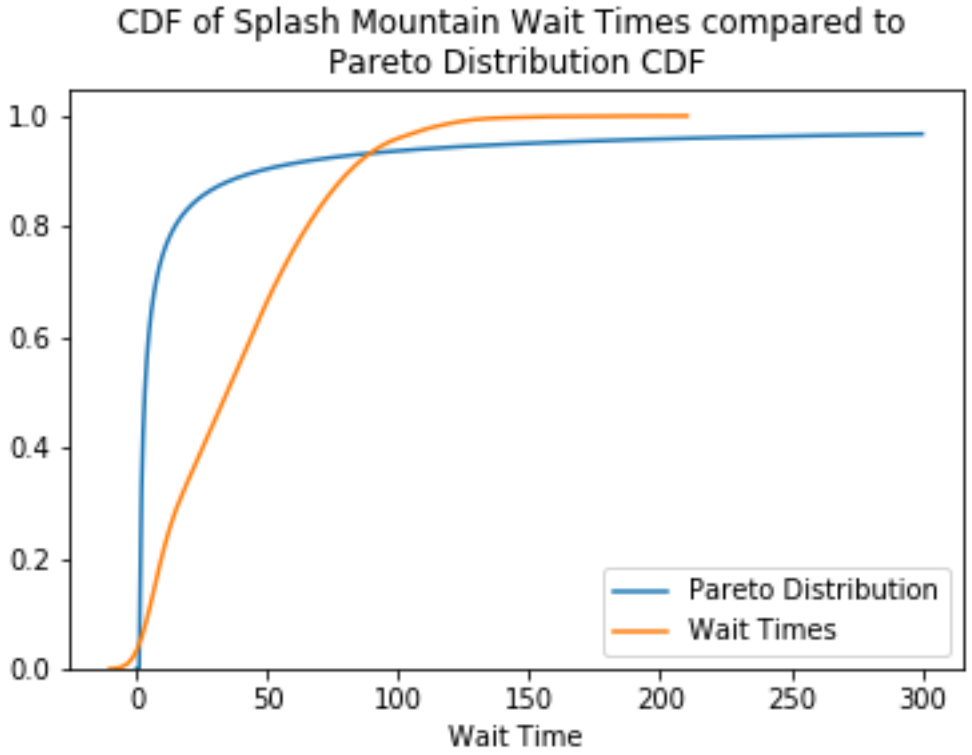


CDF



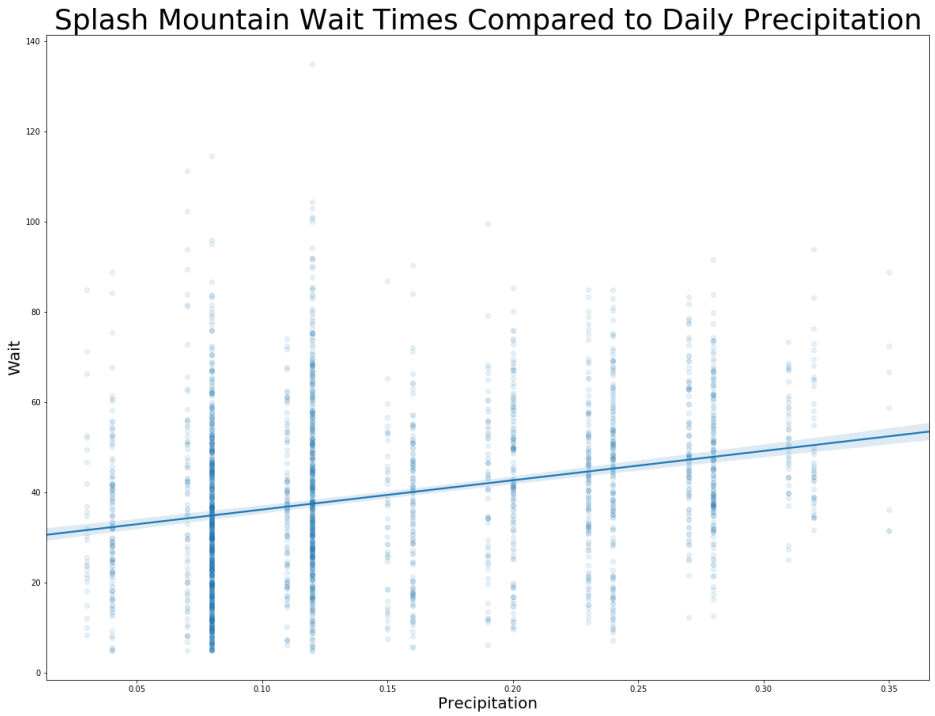
From this CDF we can see nearly all of our values exist below around the 100 minute mark. This tells us that wait times above this point are outliers.

Wait Times and the Pareto Distribution



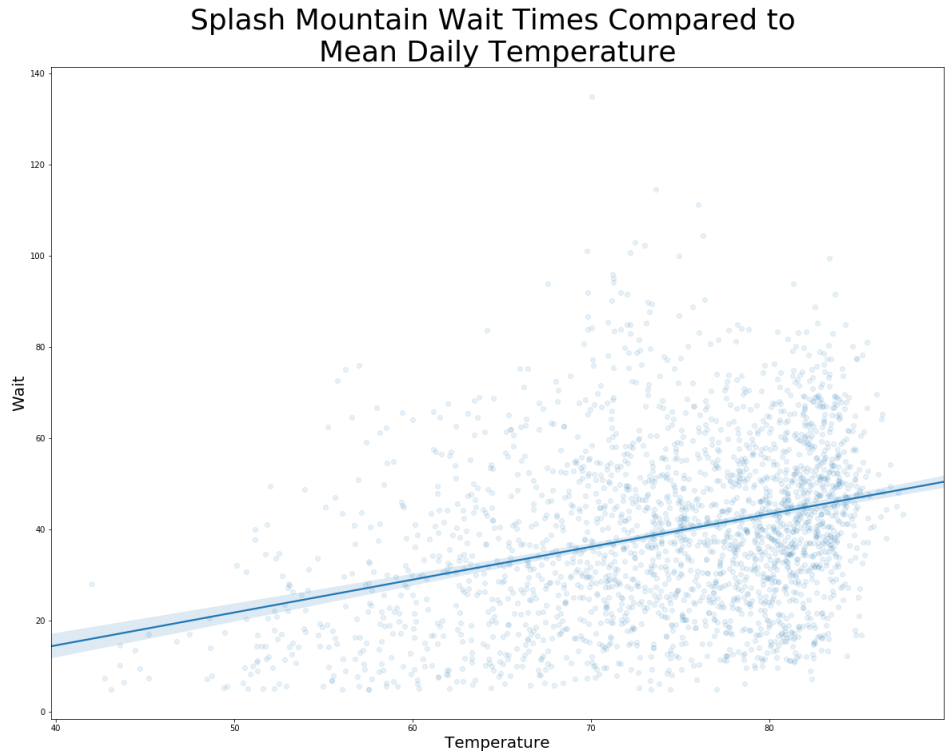
While not a perfect fit, the distribution of wait times fits the pareto distribution's idea that most of the data lies in a small portion of the distribution.

Wait Times vs. Rainfall



- Spearman R – 0.295
  - A positive relationship, however not a strong one.

Wait Times vs. Temperature



- Spearman R – 0.320
  - A slightly stronger positive relationship

Correlation Hypothesis Tests

Wait Time vs. Temperature		Wait Time vs. Rainfall	
p-value	< 0.001	p-value	< 0.001
Actual Correlation	0.321	Actual Correlation	0.271
Max Correlation Seen in Testing	0.067	Max Correlation Seen in Testing	0.074

OLS Regression

OLS Regression Results

Dep. Variable:	Wait	**R-squared:**	0.304
Model:	OLS	**Adj. R-squared:**	0.303
Method:	Least Squares	**F-statistic:**	258.6
Date:	Fri, 31 May 2019	**Prob (F-statistic):**	1.39e-184
Time:	15:21:15	**Log-Likelihood:**	-9934.0
No. Observations:	2375	**AIC:**	1.988e+04
Df Residuals:	2370	**BIC:**	1.991e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-42.0235	4.412	-9.526	0.000	-50.674	-33.373
Rain	-11.3347	5.976	-1.897	0.058	-23.054	0.384
Temp	0.6325	0.049	12.984	0.000	0.537	0.728
in_session	-10.2395	1.065	-9.617	0.000	-12.327	-8.152
park_hours	3.1204	0.177	17.636	0.000	2.773	3.467

Omnibus:	267.756	**Durbin-Watson:**	1.326
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	414.853
Skew:	0.808	**Prob(JB):**	8.24e-91
Kurtosis:	4.258	**Cond. No.**	1.41e+03

- This model uses precipitation, temperature, percentage of in-market schools in session, and park hours to explain 30.3% of the variance in mean daily wait times for Splash Mountain.
- Holiday rank was also tested, but had a large p-value (0.36) and was excluded from the final model as it did not improve the results.