Scale.ai Take-home Challenge

Submission by: Adrian Lievano
February 17, 2020

Problem:

In this challenge, candidates are asked to assess the impact of temperature on total trip count and on total trip duration. The data contains bike sharing information for different user types during a time span of nearly four years with over 9.4 million examples and 23 tracked feature columns. The feature columns contain 23 total categorical, date, and numeric variables that guide data exploration and analysis. For the purposes of this analysis, we study temperature segmented by a variety of columns (such as gender, weather events, usertype, and more).

Analysis Approach:

The analysis in the accompanying jupyter notebook is as follows:

1. **Data preparation**:
    a. During this step, we explore categorical values for categorical columns, study the distributions of each numerical values, and ensure that we are familiar with the provided data. We also sample 100,000 examples from the original dataset to make producing visualizations and insights quicker.
2. **Data Exploration**:
    a. During this step, we create summarized views of the data, aggregating temperature and total trip count data for different time intervals. We also create quick visualizations of the data to build intuition on general trends.
    b. We explore the impact of other feature columns, such as gender, usertype, or events and simplify our analysis.
3. **Statistical Analysis of Various Aggregation Strategies**:
    a. The temperature data versus trip duration contains many data points that make understanding the trend difficult. We grouped temperature data on various filters (year, month, week, hourly time group) and (year, month, week group) and by different statistics of the temperate (mean or max value) at the respective time point.
    b. After building a variety of different data frames that exploit different aggregation techniques, we build a 95% confidence interval from the null case (the original, unaggregated temperature distribution) to determine if the new aggregations temperature distributions fall outside of the confidence interval. If they did, we would determine that these distributions are not representative of the null temperature distribution.
4. **Regression Analysis**:

a. Linear regression models are trained to determine the slope and y-intercept of the temperature versus trip duration variables using the two different aggregated temperature distributions. We score the fit and determine an approximate relationship between temperature and trip duration.

Results:

1. **Warmer Temperature is associated with an increase in total trip duration and total trip counts**.
    a. Average trip duration increases with temperature; it is highly correlated (0.48) with trip count and trip duration.
    b. The total number of bike trips tends to increase with warmer temperatures up to about 80 degrees fahrenheit. Afterward, the number of total bike rides decreases.
2. **Gender has little impact on trip duration and total trip volume**.
    a. After segmenting for gender (male, female), the results tend to be the same.
3. **Usertype**:
    a. Because nearly 99.99% of user types were subscribers, as opposed to customers, we reframed from making any conclusions from this data given the degree of imbalance in the classes.
4. **Weather**:
    a. 88.8%, 5.5%, 4.5%, and 1.5% of the 100,000 samples of bike sharing data correspond to weather related events of cloudy, clear, rain or snow, and not clear or other. A majority of total trips occur on cloudy days and very few during thunderstorms or during rain or snow. Weather events, however, do not have a significant impact on trip duration.
5. **Regression:**
    a. An equation with an R2-score of 0.31 is generated.

Supporting Figures:

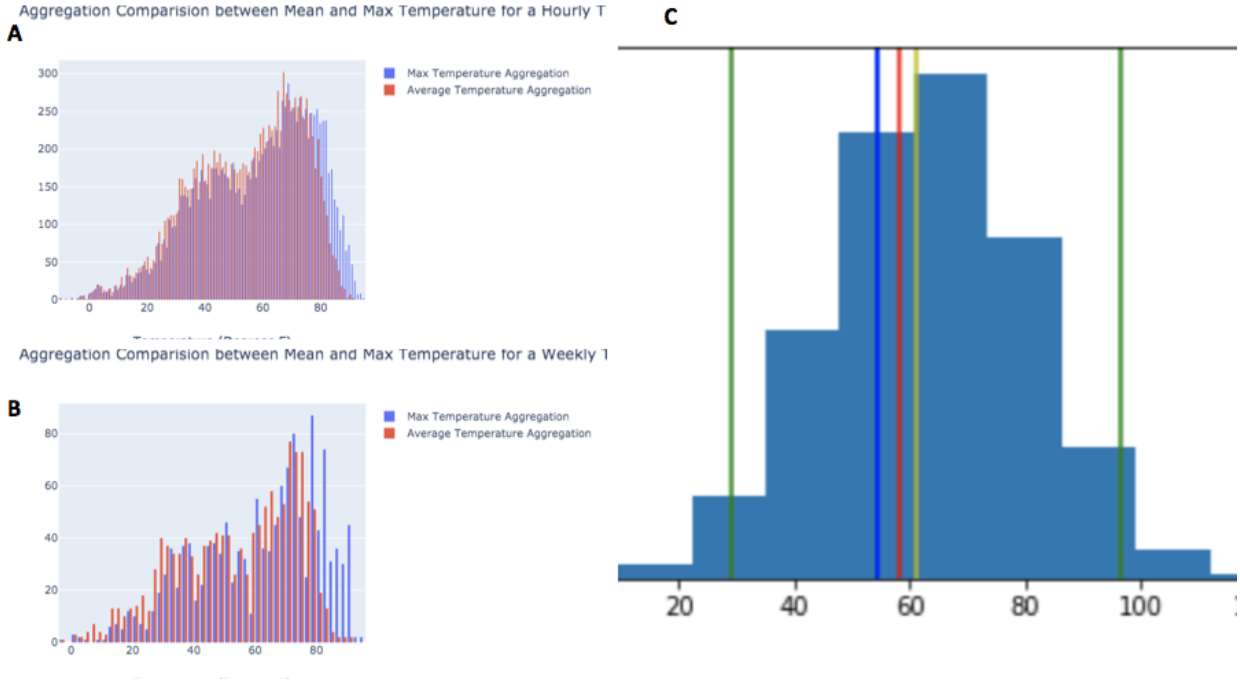

Figure. Sample distributions of mean and max temperature based on weekly and hourly time aggregations on the bike-share data. (A) Hourly aggregation for average and maximum temperature distribution. (B) Weekly aggregation for average and maximum temperature distribution. (C) A simulated temperature distribution from the original dataset. The green lines represent a 95% confidence interval. The blue line represents a weekly aggregation. The red represents an hourly aggregation on the original data. The three sample distributions are statistically insignificant and therefore we can safely use a hourly or weekly time-based average temperature aggregation.

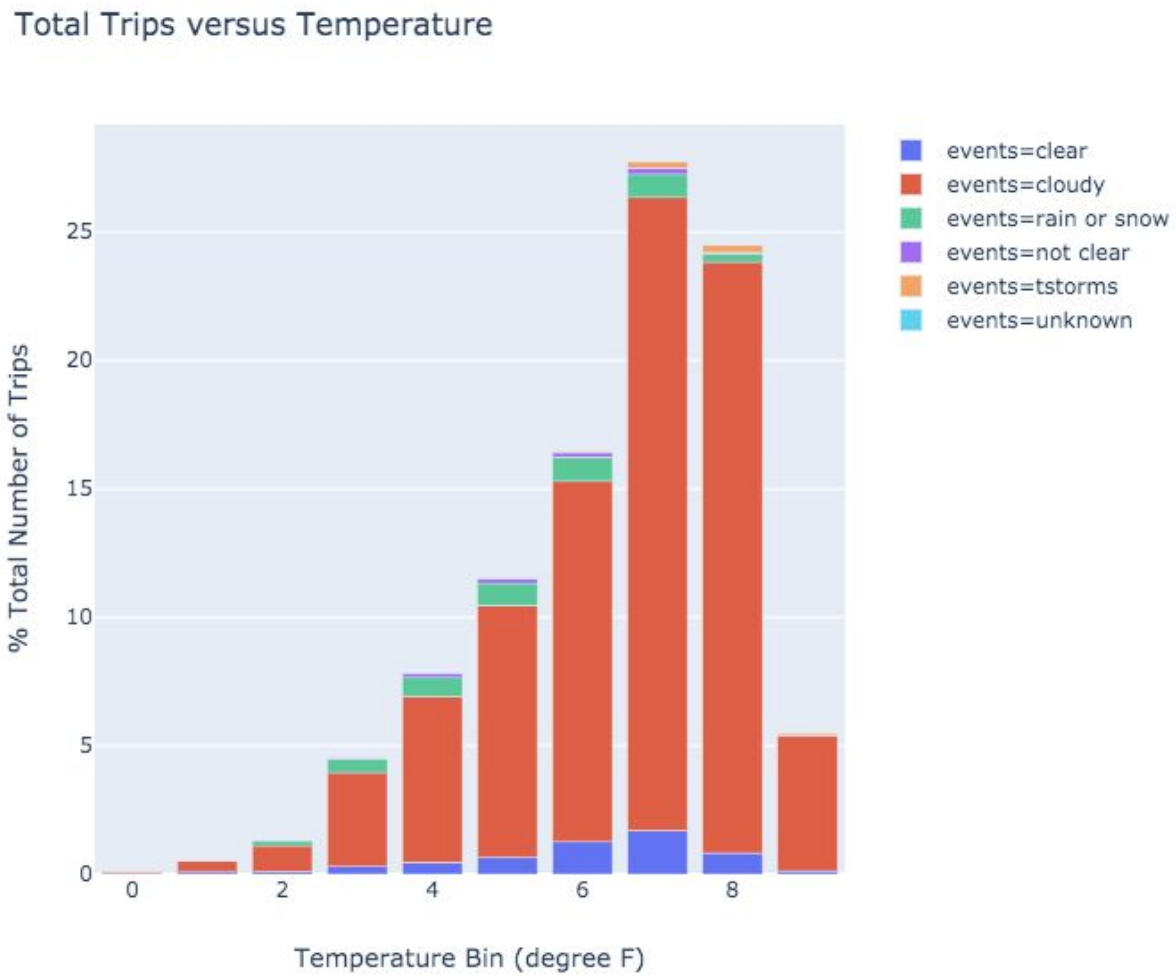

Figure. Temperature versus % Total Number of Trips from 100,000 bike trip examples segmented by weather events.

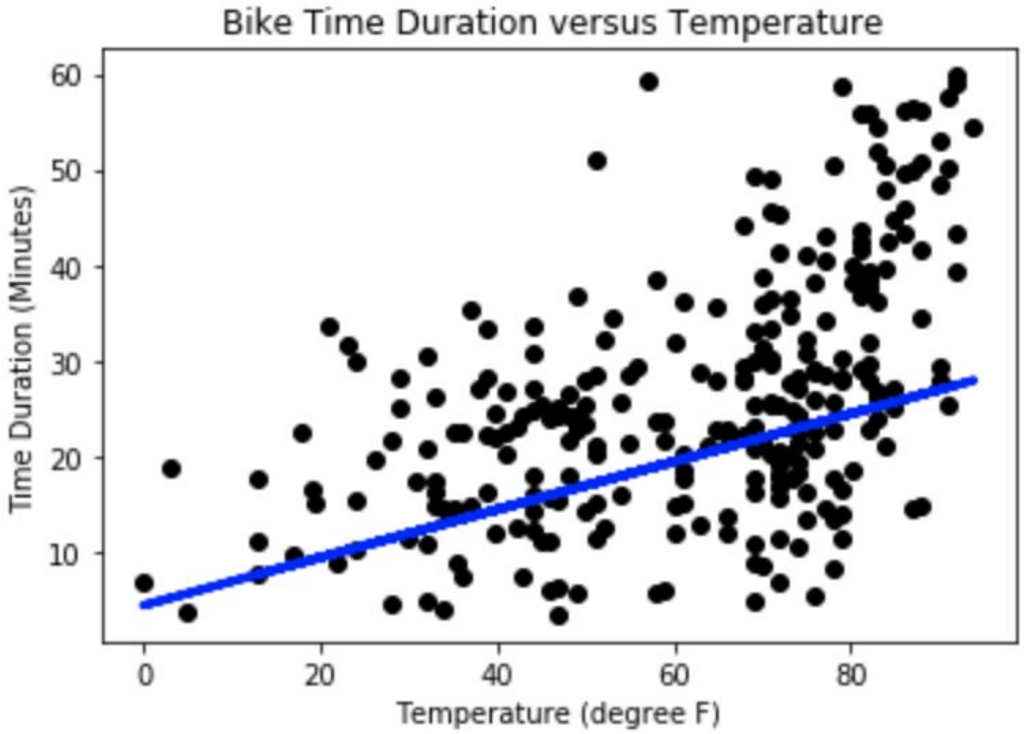

Figure. A linear regression fit between temperature (degree fahrenheit) and total trip duration (minutes). R2-score of 0.310. The line's slope is 0.31. The y-intercept is 7.66 minutes.

Jupyter Notebook & Github Repository:

https://github.com/adrianlievano/scale-ai-challenge