Bike Sharing in Mexico City

Final Project Introduction to Data Science

Jesus L. Trujillo

October 2014

The project

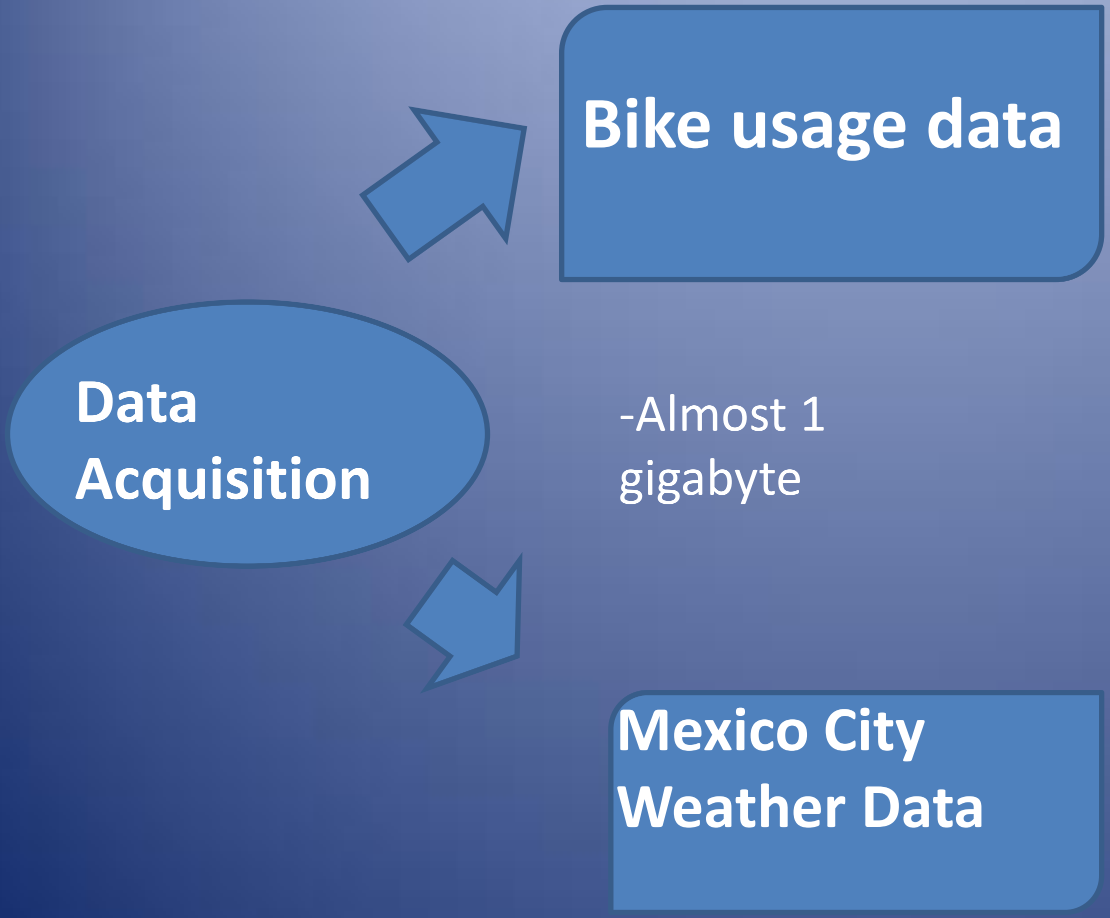
- Bike sharing system information from Mexico City ‘Ecobici’: http://datos.labplc.mx/datasets/view/ecobici
- The program started in February of 2010 and is considered one of the most successful in the world:

http://www.economist.com/node/16591116

What do we know about bikes?

- To model this problem I followed the Kaggle competition structure:
  - Predict demand of bikes using weather data and characteristics of users
  - Data already processed but offered little insights on what else is going with bike data

Process



Data Acquisition

Bike usage data

3 different files:

-User

-Trips

-Stations

-Almost 1 gigabyte

Mexico City Weather Data

365 different ‘files’:

-Web scrapping process to gather all of the data

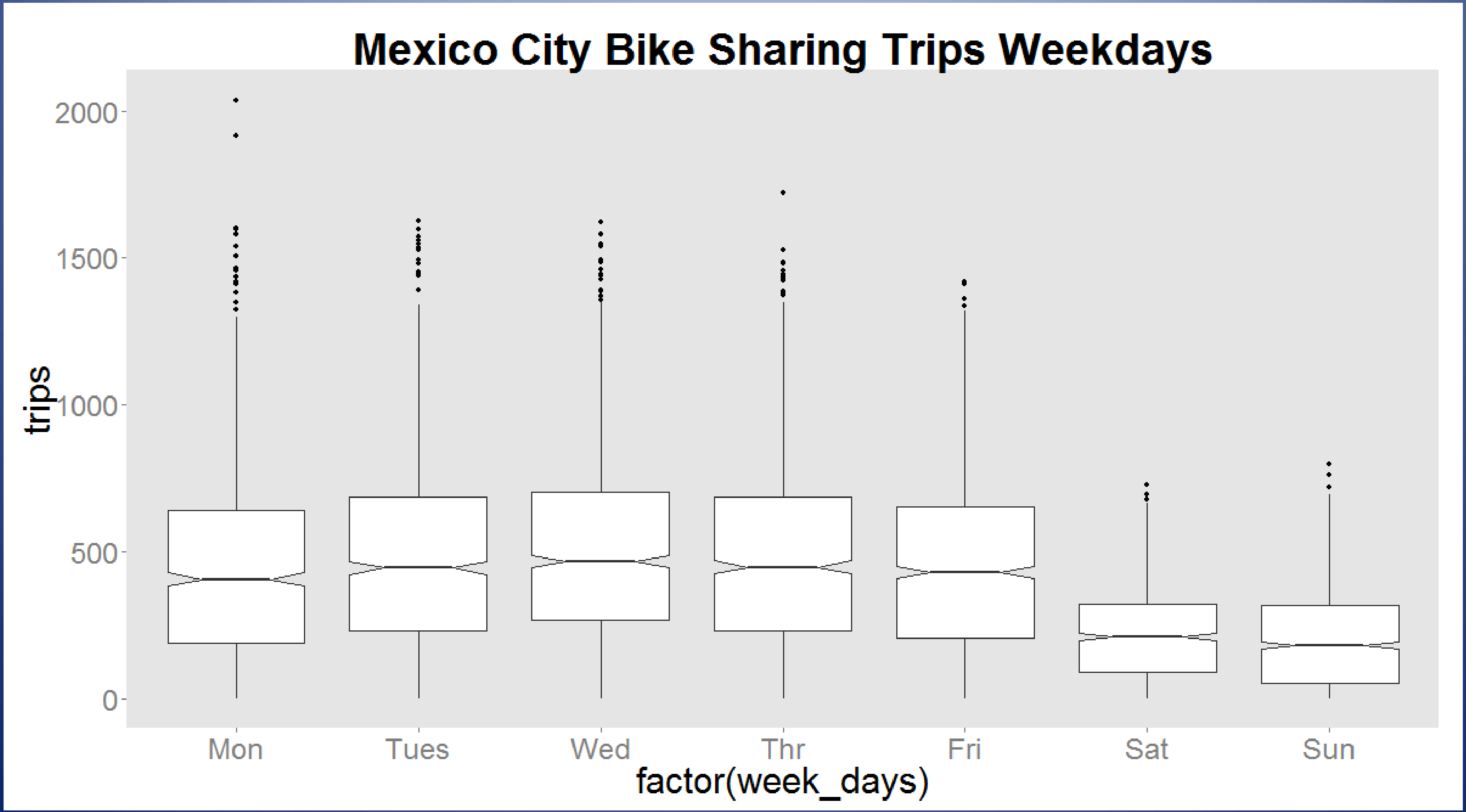
Two main objectives

- Descriptive statistics:
  - Number of trips
  - Average duration and distance traveled
  - Hourly, Daily and Monthly usage etc.
- Predictive modeling:
  - Boosted Regression Model
  - Bayesian Ridge Regression
  - Support Vector Machine Regression

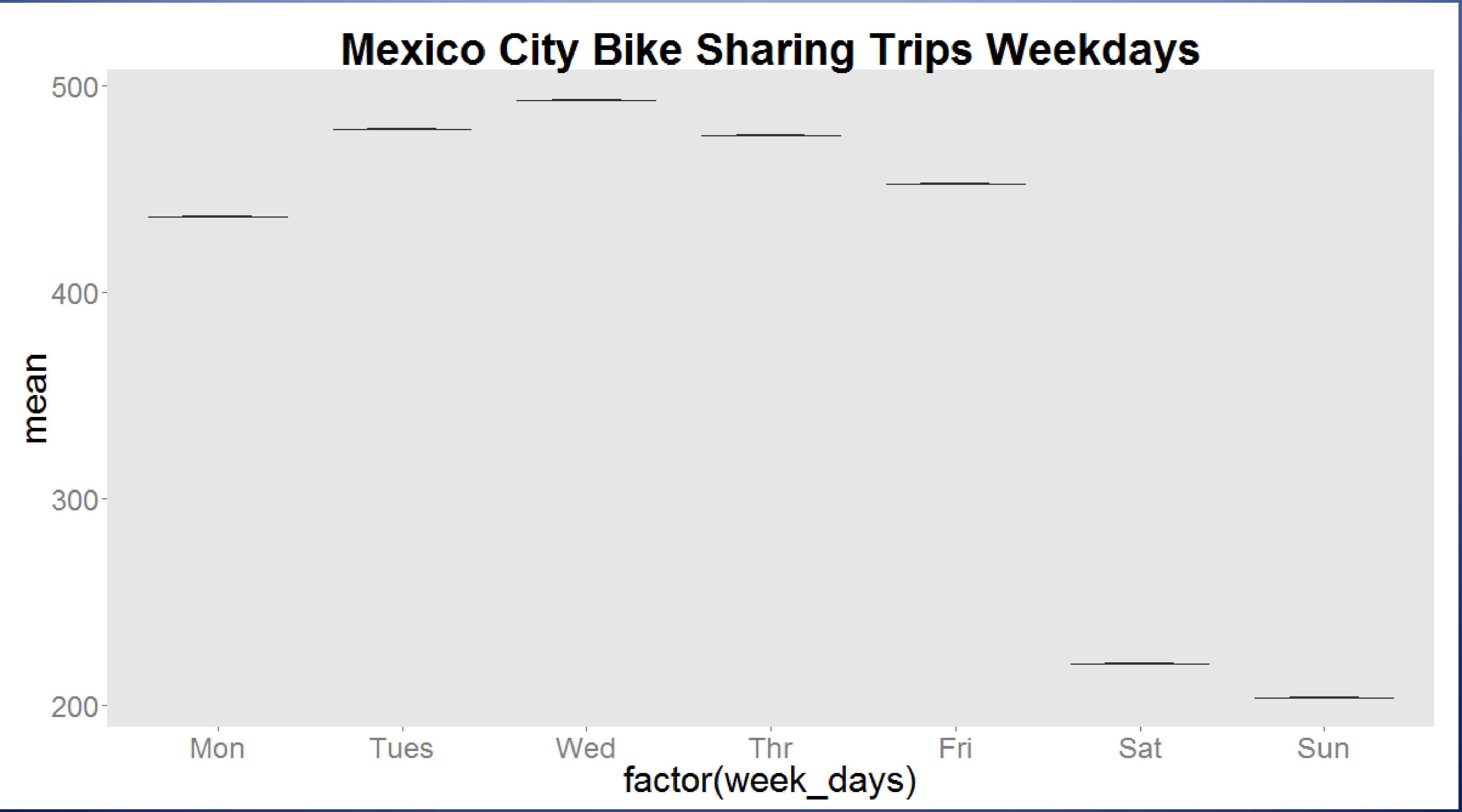
Descriptive statistics

- In 2012 there were 2,874,749 travels, or almost 8,000 trips a day
  - – The average duration of a trip is: 14.8 minutes
  - – And the average distance traveled per trip is: 1.13 km.

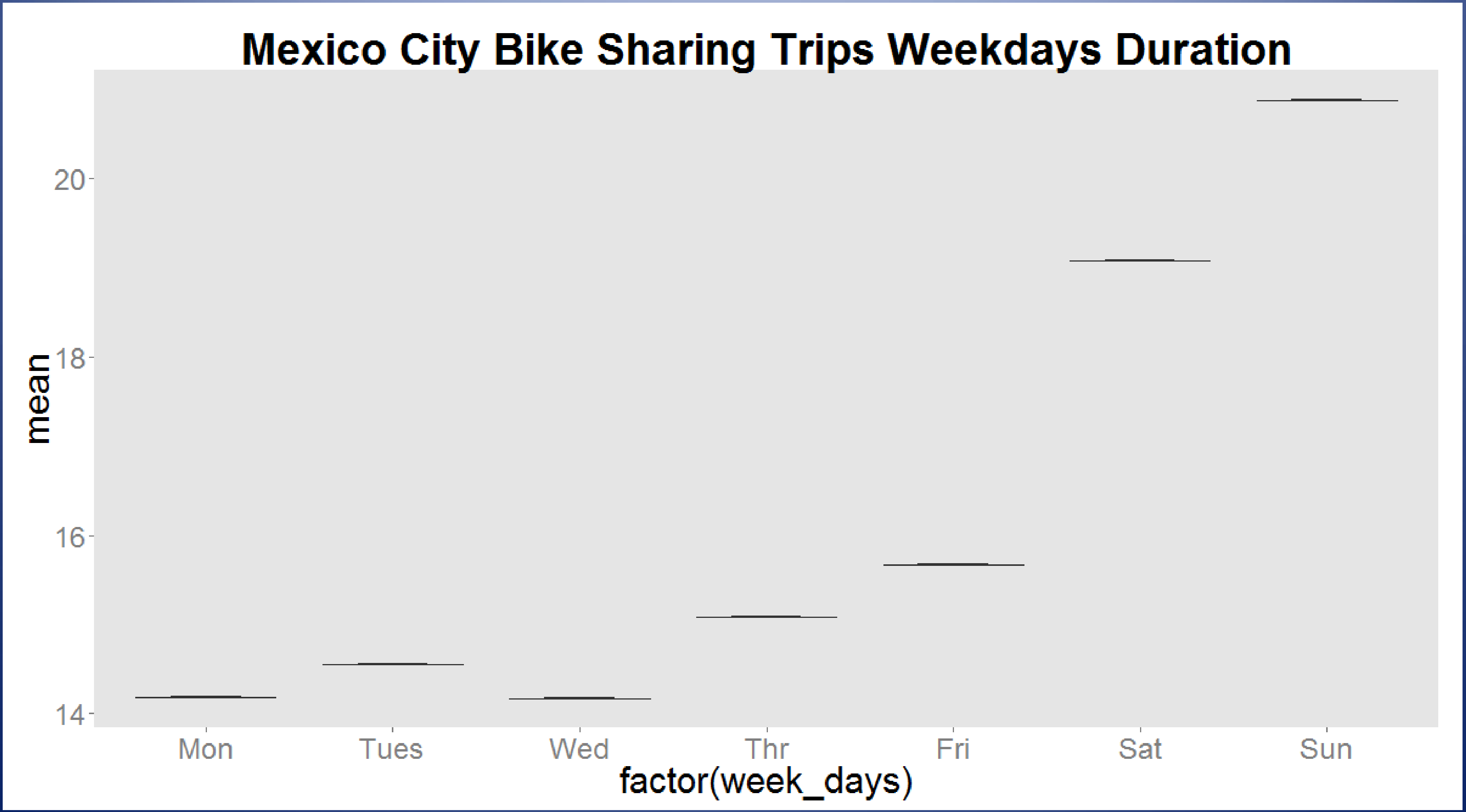
Usage by Day of the Week



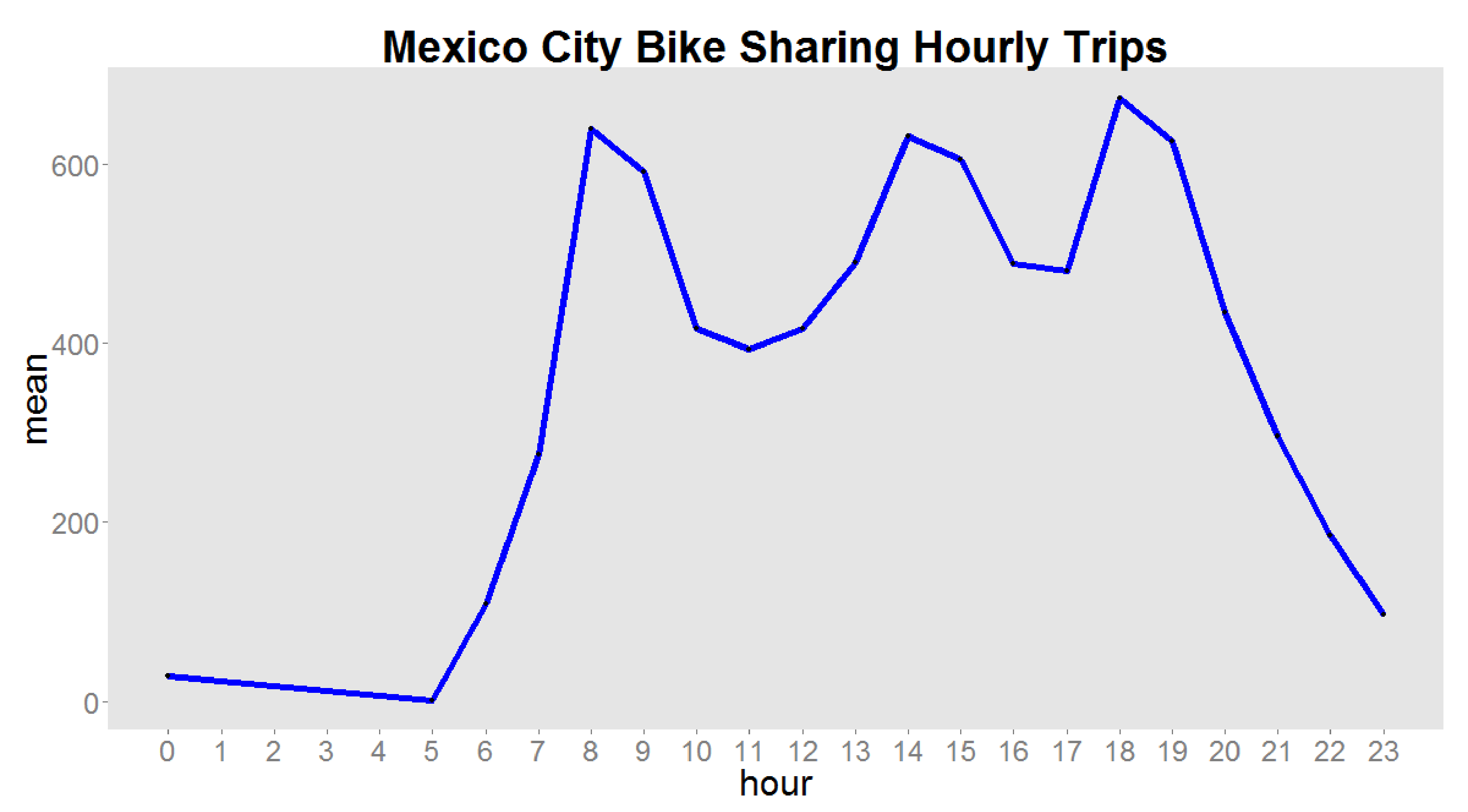
A more accurate description of usage by day of the week



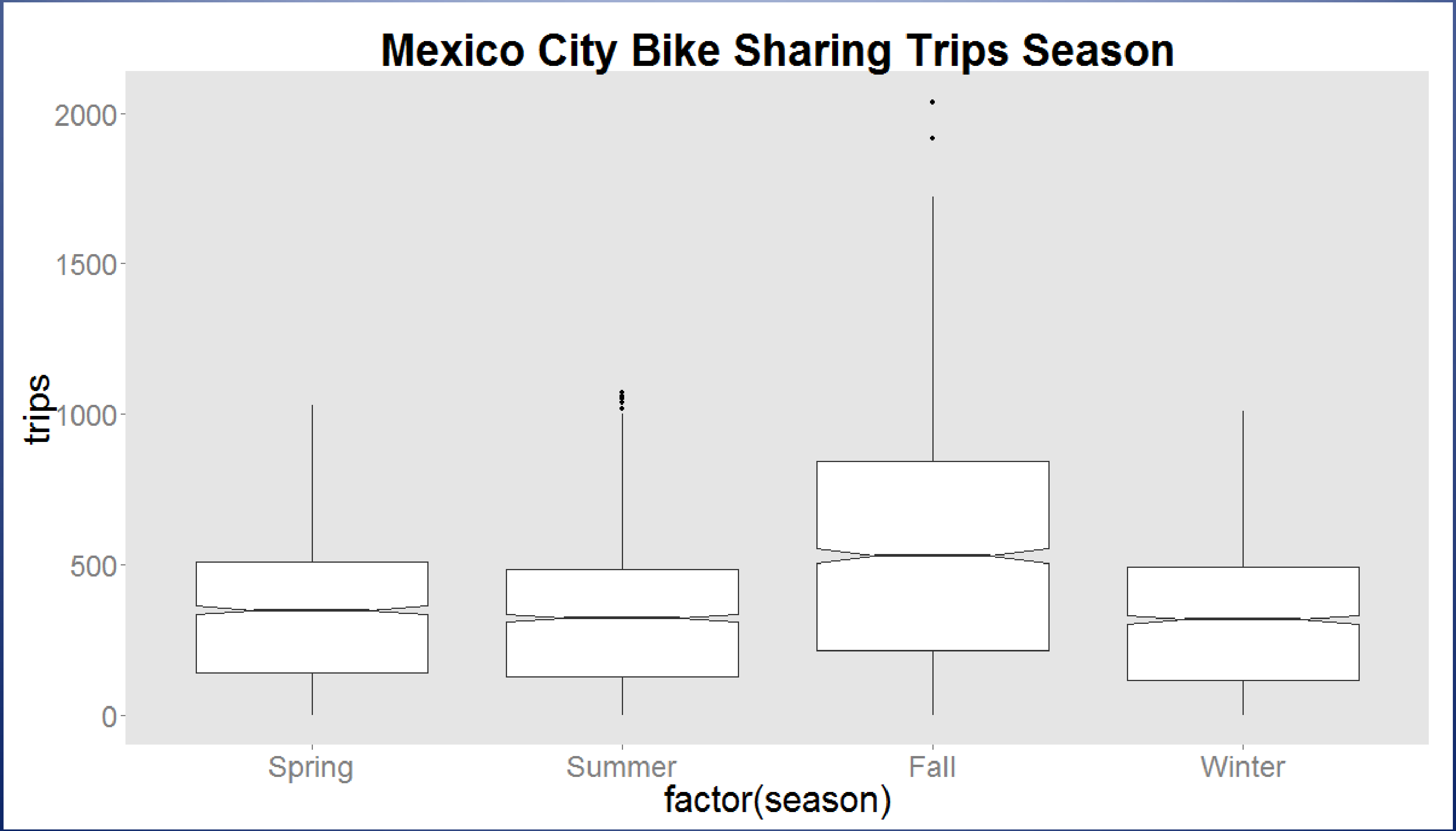
But if time matters…

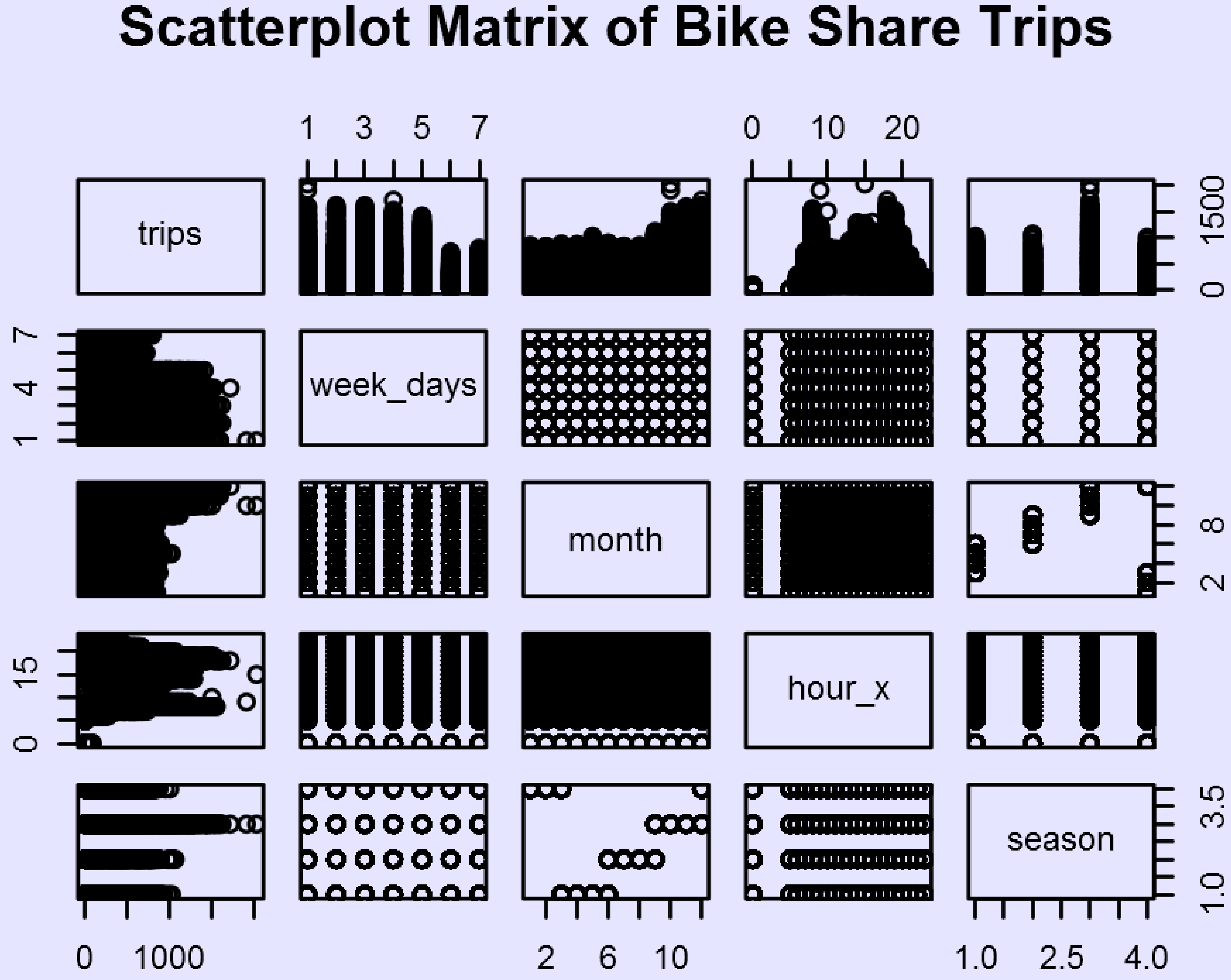


What about hourly?



What about seasons?



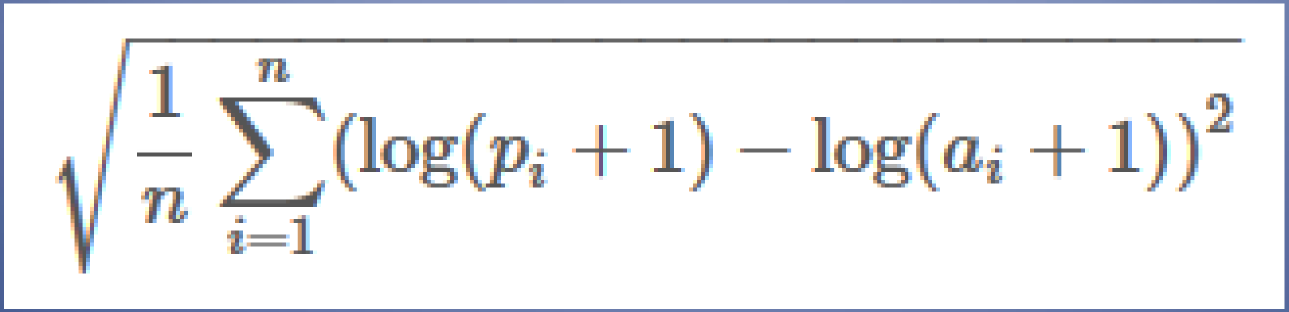

Scatterplot Matrix of Bike Share Trips
trips
week_days
month
hour_x
season
1
3
5
7
0
10
20
0
1500
2
8
1.0
3.5
7
4
1
15
0
0
1000
2
6
10
1.0
2.5
4.0

Predicting Bike Demand

- After transforming the data I ended up with a data frame composed of 7,320 observations and 13 predictors (X) and a target (Y):
  - The predictors are:
    - week_days, month, hour, season, Gust Speed, Humidity, WindirDegrees, Conditions, Events, Dew Point, TemperatureF, VisibilityMPH, and WindSpeed
  - The target variable is:
    - Number of trips in a given hour

The Kaggle Method

- Accuracy is evaluated using the Root Mean Squared Logarithmic Error (RMSLE).



$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

- *n* is the number of observations in the test set
- *pi* is your predicted count
- *ai* is the actual count
- log(*x*) is the natural logarithm

Boosted Regression Model

- First step in the model was to normalize the data and make sure that there were no missing values:
  - Scikit Learn really does not like NaN
- The first model specifications (default) gave a pretty mediocre prediction rate:
  - 10.4313000472
  - Actual benchmark: 0.24976
- Optimized model gives:
  - 6.34154256614

Bayessian Ridge Regression

- Did a better job predicting in variation:
  - Predicted values ranged from 0 to 987
    - Max value in my dataset is 1687
- The R squared was a bit better although I could not really improve accuracy at all:
  - rmse=6.4

Support Vector Machine Regression.

- This was the model with the highest success rate:
  - rsme= 6.29587856774
- Did better with data variation than Gradient boosting regression.

Next steps

- Add more variables to the dataset:
  - If the weekday is a holiday
  - Better temperature data
  - Pay day
- Better transformation of the categorical variables:
  - Many of the numeric values are meaningless
- Understand a bit better what is going in the blackbox
- Find clean curated data!