Bike Sharing in Washington D.C.

Ashley O'Mahony
ashleyomahony.com | March 2019

Context

2011

1,500 bicycles

165 stations

18,000 members

2012

1,650 bicycles

175 stations

22,200 members

Objectives

1. Predict the amount of users on an hourly basis
2. Ensure high level of service and availability
3. Optimize Logistics and Maintenance Teams

1.
Project Structure

Project Organization

Model and Predictions

3 Using a Linear Regression algorithm, test the impact of the features on the model score (R^2)

Data Preparation and Features Construction

2 Based on Exploratory Data Analysis and Machine Learning principles

GitHub + GitKraken

1 Teamwork improved using collaborative developer tools

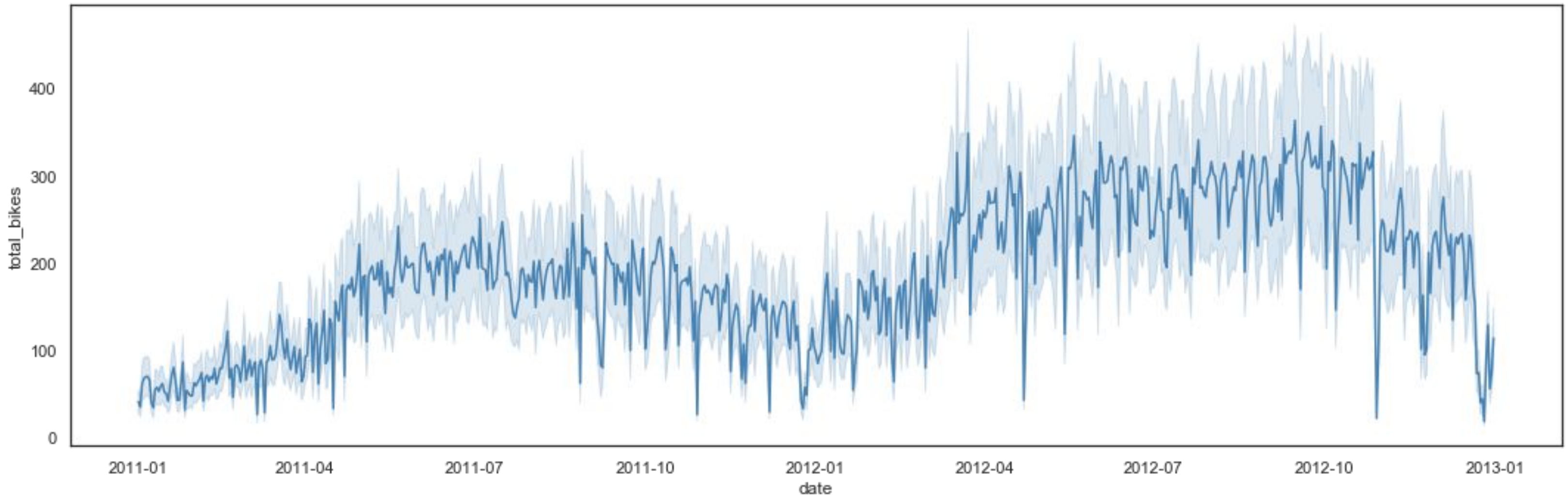
Machine Learning Process

01	EDA and Data Preparation	• Remove *Casual*, *Registered*, *Holiday*, *Feeling Temperature* • Scaling, Skewness, Encoding
02	Machine Learning Strategy	• Train set: Jan 2011 - Jul 2012 • Test set: Aug 2012 - Dec 2012 • Time Series Cross Validation (10 folds)
03	Feature Engineering	• Patterns on Dates and Hours • Peak Detection • Exceptional Weather Conditions • Polynomials
04	Selection and Final Metric	• Recursive Feature Elimination • Manual Selection • Model Predictions vs Reality

2.
Data Exploration Key Insights

2011-2012 Utilization

Our bike sharing system gets more users every year.



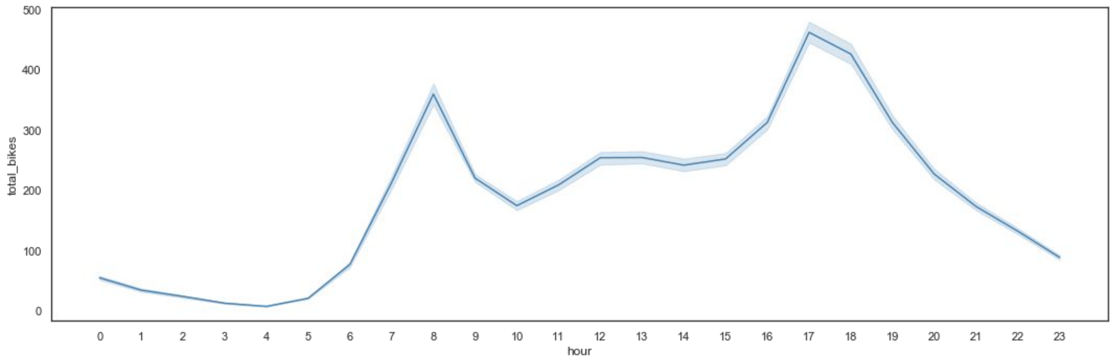
Number of bikes used over 2011-2012

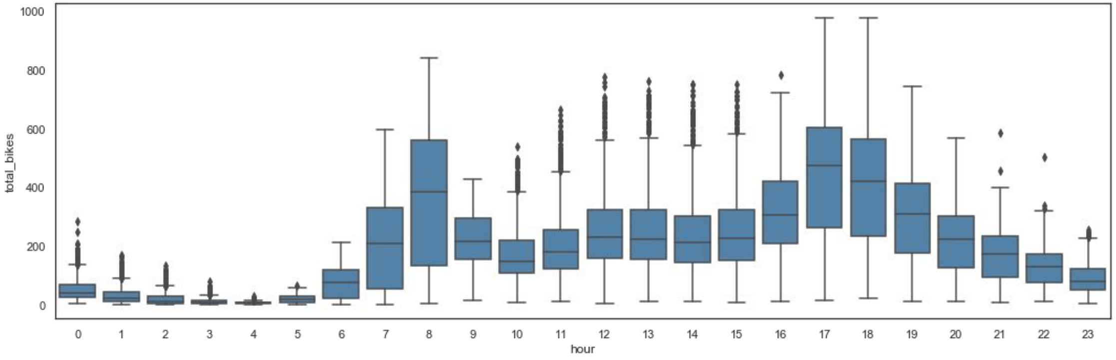
Utilization by Hour

- Day time usage

- One peak around 8am
- One peak between 5-6pm

- Up to 1000 bikes within an hour

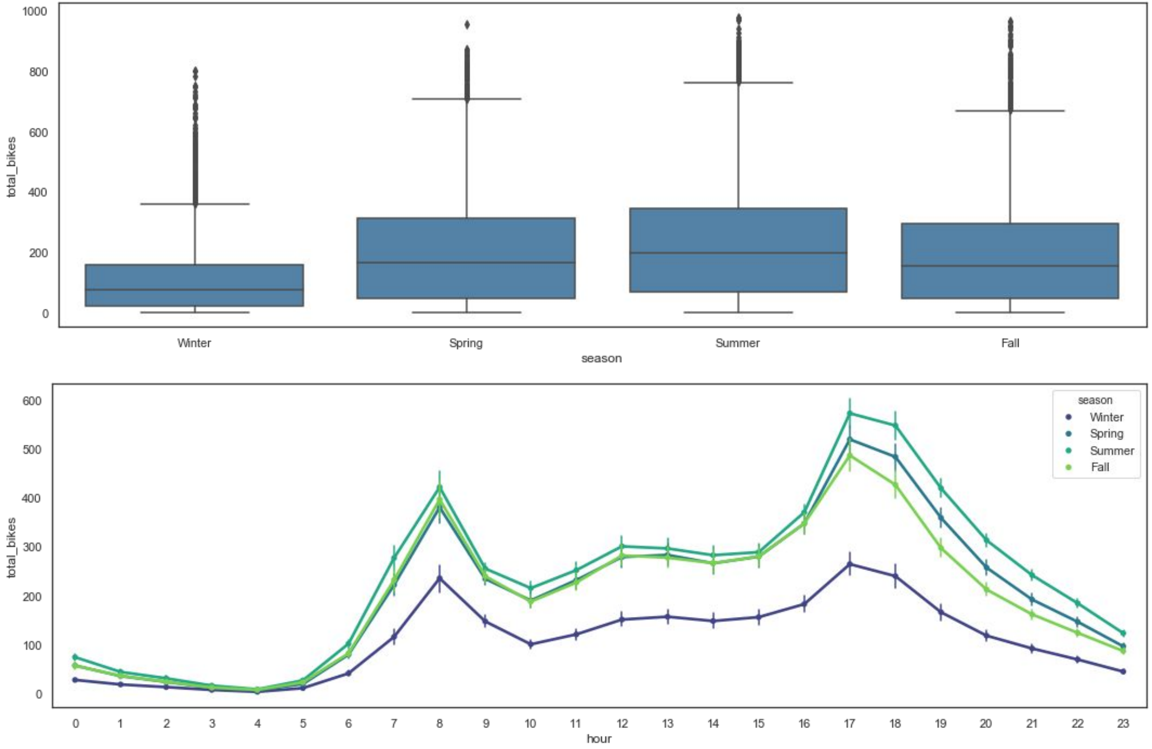




Number of bikes used by Hour

Utilization by Season

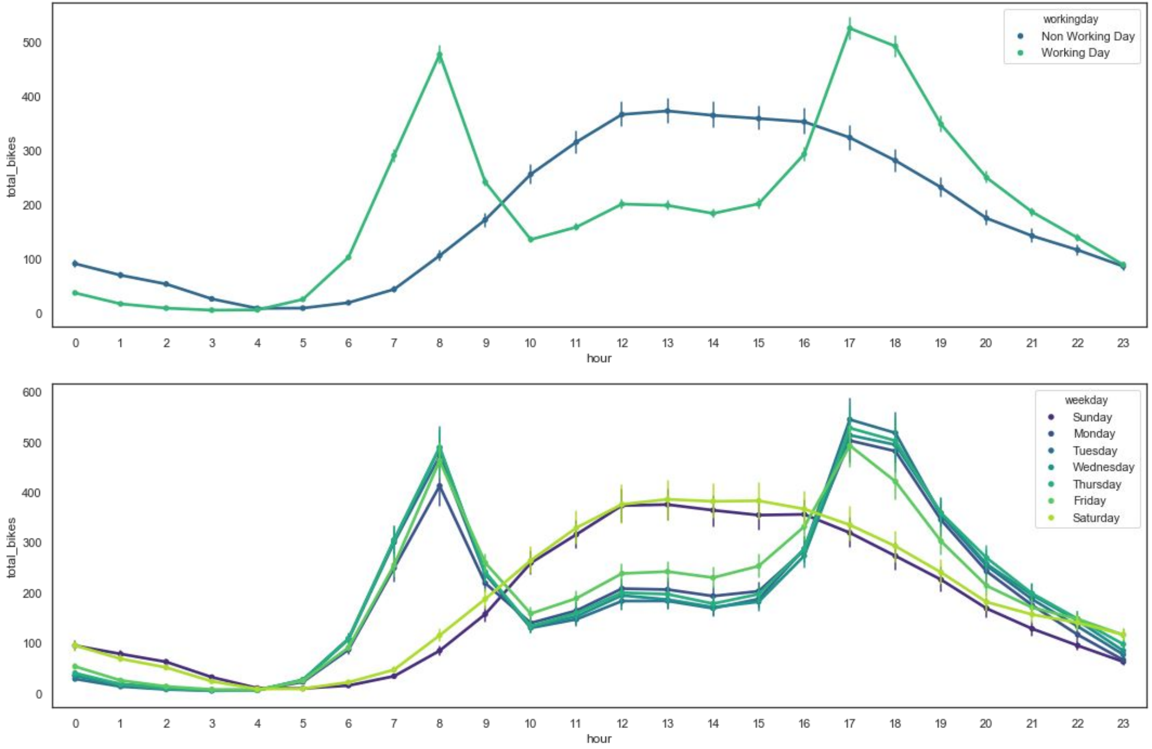
- **Summer is the high season**
- **Winter is the low season**
- **Spring and Fall have similar utilization shapes**



Overall and Hourly Utilizations by Season

Working Days

- **2 peaks on working days during commuting hours**
- **No peak during non working days, but higher overall utilization in the afternoon**
- **Slight change of shape on Fridays, maybe because people leave work earlier on that day**

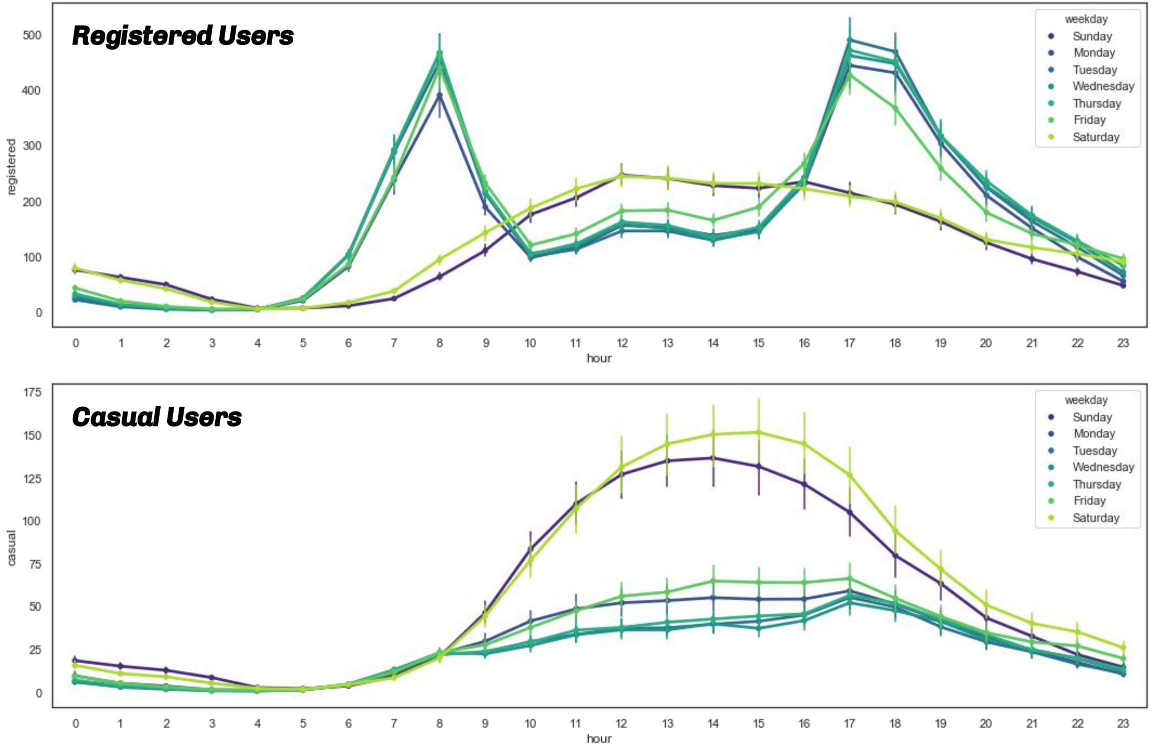


Hourly Utilization on Working/Non Working Days

Working Days

- **Clear difference in behaviours between our registered users and the casual users**

- **Commuting and Leisure effects**

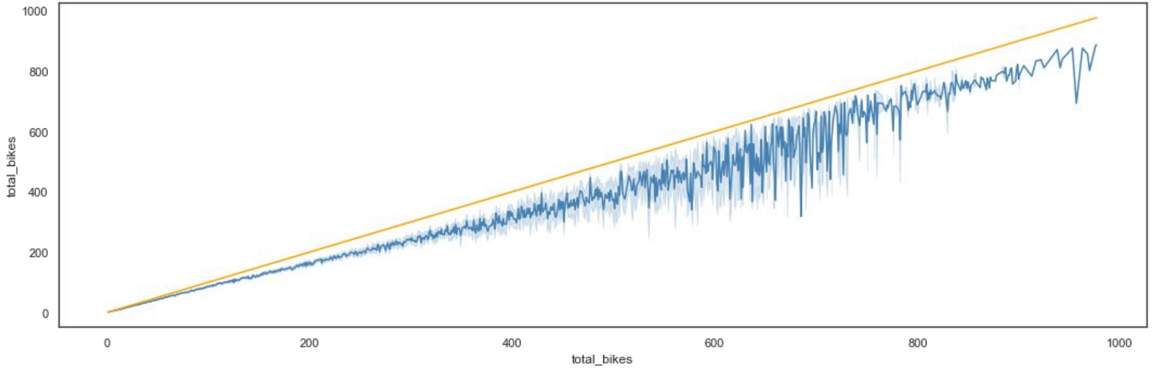


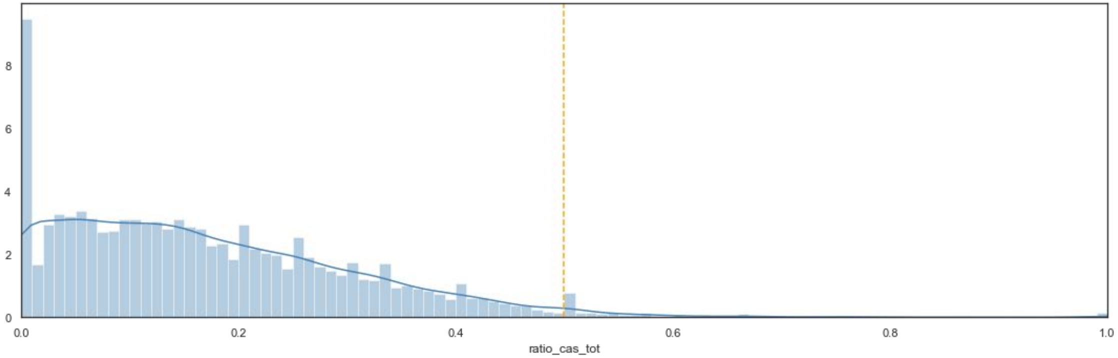
Hourly Utilization on Working/Non Working Days

Utilization by User Type

- **Most users are registered**
- **High correlation with the Total number of bikes**

→ ***Casual* and *Registered* users information removed from the dataset**

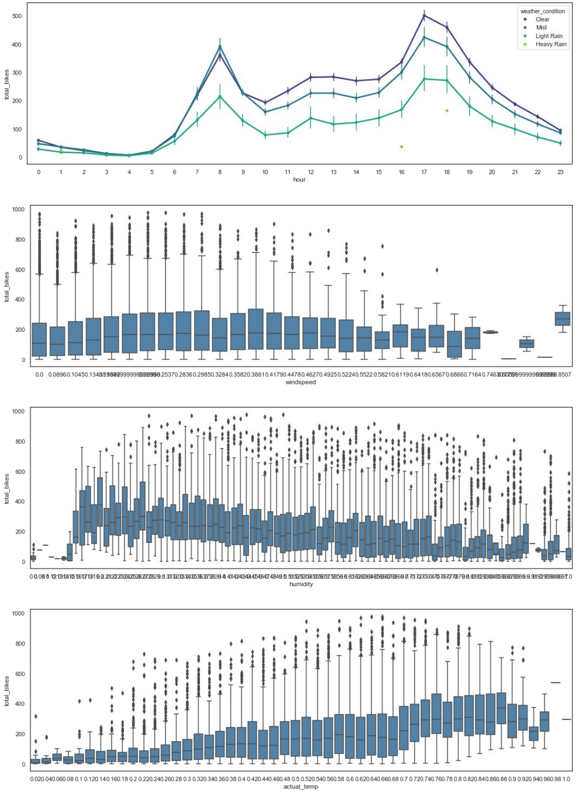




Ratio of Registered Users

Weather Conditions

- **Weather conditions have a small impact on the service utilization**
- **Rain has the clearest effect**
- **Strong Wind discourages users**
- **Humidity and Temperature seem to have less influence**



Rain

Wind

Humidity

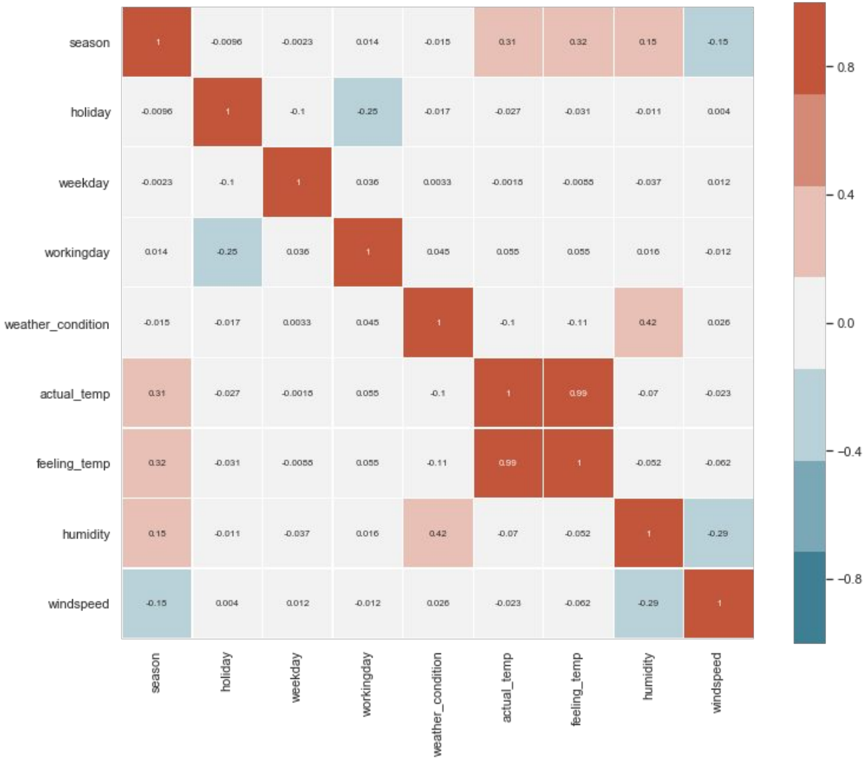
Temperature

Utilization based on Weather Conditions

Correlations

- **Correlation between Actual and Feeling Temperatures is clear**

- **No other strong correlation between other variables**



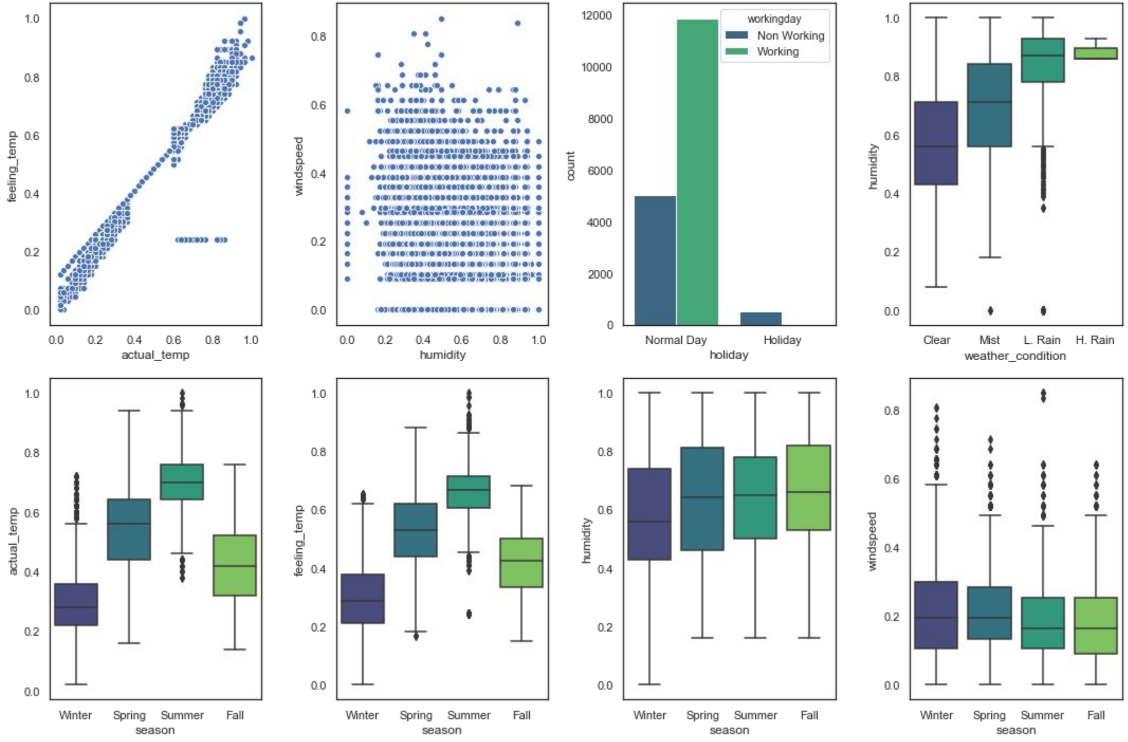
Correlation Matrix

Correlations

- **Actual and Feeling Temperatures plot is clear**

- **Every Holiday is a Non-Working Day**

→ *Feeling Temperature* and *Holiday* information removed from the dataset

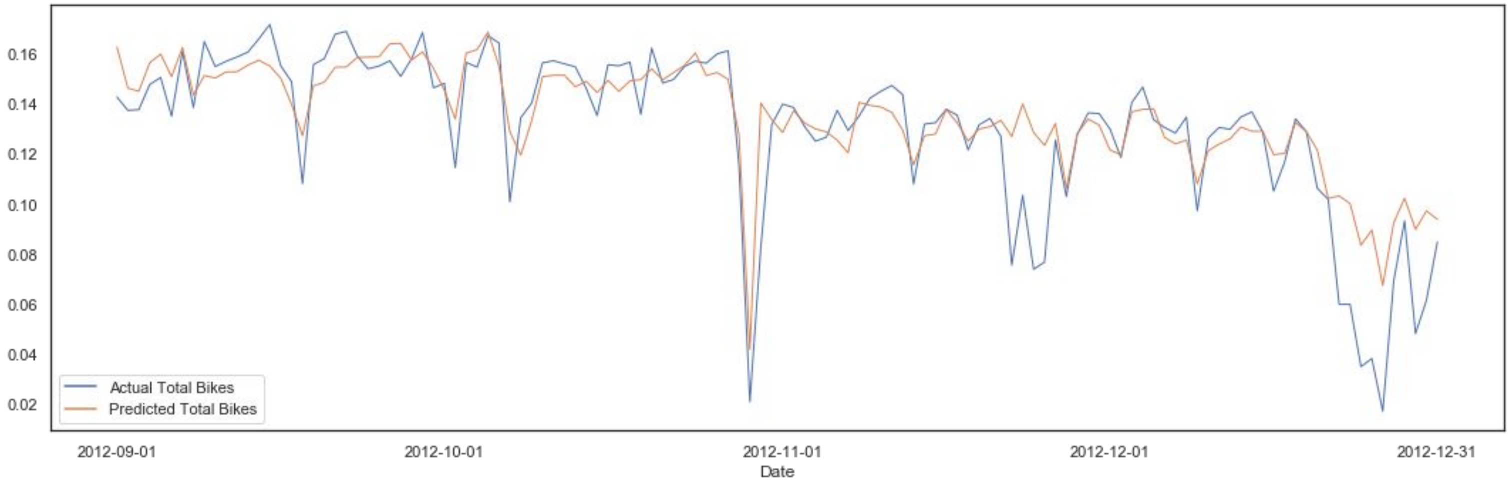


Pair Plots

3.
Features Construction

Baseline

Features: 57 **R^2: 0.76**



Baseline Predictions vs Reality

Reminder - Features removed from dataset:

Casual, Registered, Holiday, Feeling Temperature

Calendar Features

R^2

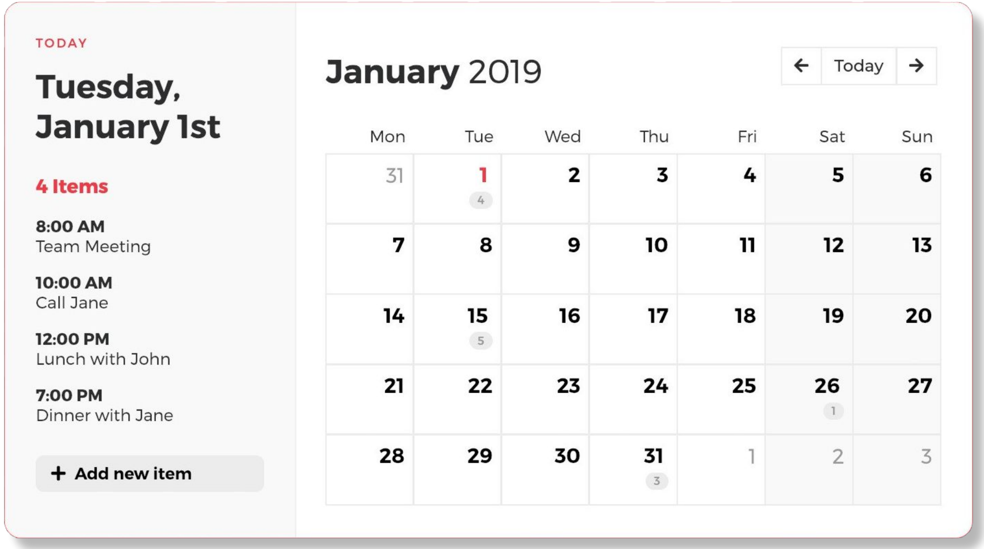
0.76 — **0.75**

BASELINE

Without Casual, Registered, Holiday, Feeling Temperature

DAY AND MONTH-DAY

Add Day and Month-Day to identify patterns specific to special dates



Peaks

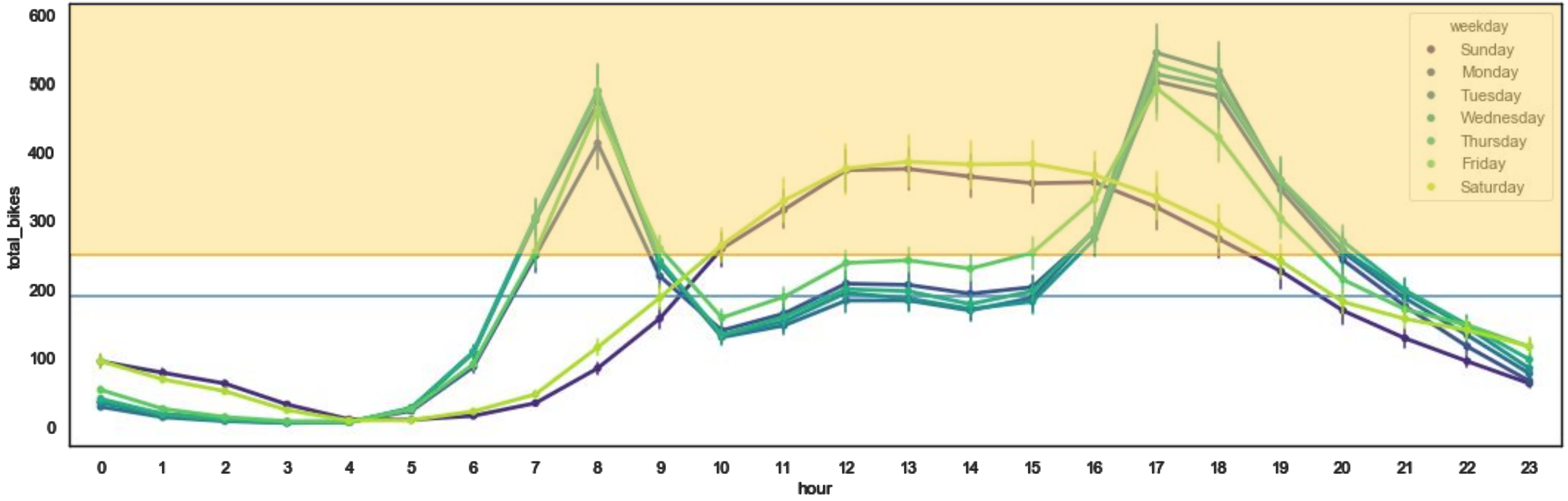
R^2

0.76	0.75	
BASELINE	**DAY AND MONTH-DAY**	**PEAKS DETECTION**
Without Casual, Registered, Holiday, Feeling Temperature	Add Day and Month-Day to identify patterns specific to special dates	Flag hours with utilization higher than a dynamic threshold

Peaks

(1+ x%)

Mean Total Bikes



Peaks

R^2

0.76	0.75	0.87
BASELINE	**DAY AND MONTH-DAY**	**PEAKS DETECTION**
Without Casual, Registered, Holiday, Feeling Temperature	Add Day and Month-Day to identify patterns specific to special dates	Flag hours with utilization higher than a dynamic threshold

Weather

R^2

	Baseline	Day and Month-Day	Peaks Detection	Exceptional Weather
R^2	0.76	0.75	0.87	0.87

BASELINE

Without Casual, Registered, Holiday, Feeling Temperature

DAY AND MONTH-DAY

Add Day and Month-Day to identify patterns specific to special dates

PEAKS DETECTION

Flag hours with utilization higher than a dynamic threshold

EXCEPTIONAL WEATHER

Use the difference with the Season or Month averages for Humidity and Temperature

Polynomials

R^2

0.76	0.75	0.87	0.87	0.87
BASELINE	**DAY AND MONTH-DAY**	**PEAKS DETECTION**	**EXCEPTIONAL WEATHER**	**POLYNOMIALS**
Without Casual, Registered, Holiday, Feeling Temperature	Add Day and Month-Day to identify patterns specific to special dates	Flag hours with utilization higher than a dynamic threshold	Use the difference with the Season or Month averages for Humidity and Temperature	Use polynomials of numerical features

Hour Bins

	BASELINE	DAY AND MONTH-DAY	PEAKS DETECTION	EXCEPTIONAL WEATHER	POLYNOMIALS	HOUR BINS
Feat.	57		58			40
R^2	0.76	0.75	0.87	0.87	0.87	0.83
	Without Casual, Registered, Holiday, Feeling Temperature	Add Day and Month-Day to identify patterns specific to special dates	Flag hours with utilization higher than a dynamic threshold	Use the difference with the Season or Month averages for Humidity and Temperature	Use polynomials of numerical features	Reduce the number of hour variables by binning

4.
Model Selection

RFE

BASELINE	PEAKS DETECTION	HOUR BINS
R^2 0.76	R^2 0.87	R^2 0.83
57 Features	58 Features	40 Features
	RFE	RFE
	R^2 0.86	R^2 0.82
	54 Features	36 Features

4 Features Eliminated:

Humidity | Actual Temperature | Wind Speed | Working Day

Manual Feat. Selection

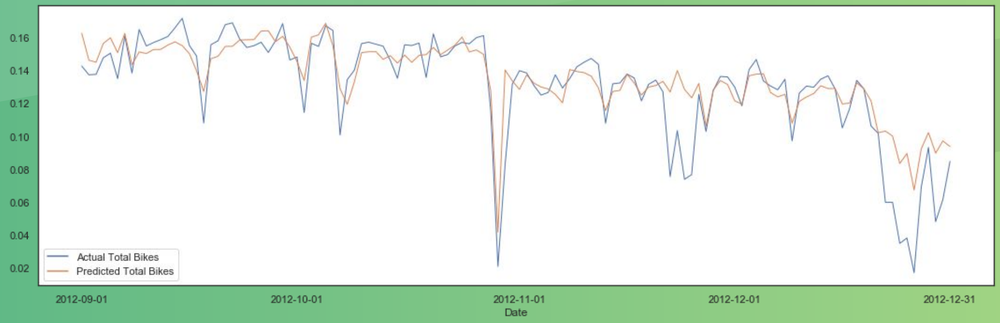
Features Kept	Features Removed
Year	Actual Temperature
Month	Humidity
Days of the Week	Windspeed
Hours	Weather Condition
Peak Detection	Working Day Flag
	Seasons

Features: 46 **R^2: 0.85**

BASELINE

Features: 57 R^2: 0.76

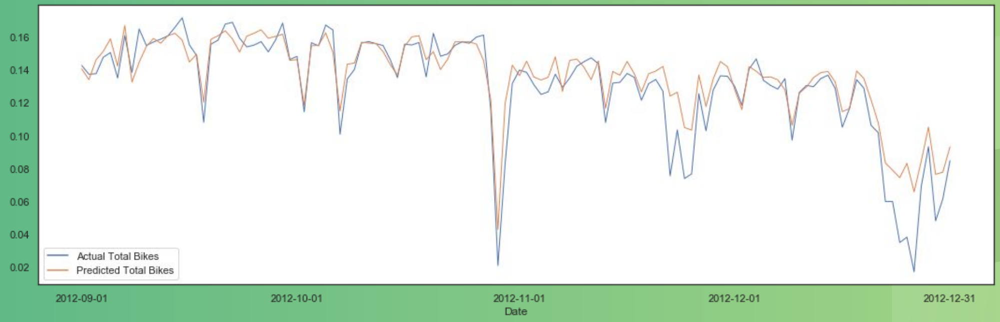
Risk of shortage during peaks



PEAKS DETECTION

Features: 54 R^2: 0.86

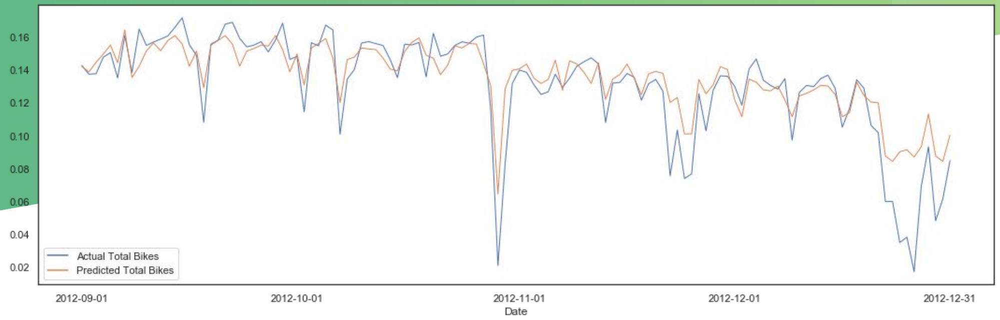
Better peaks anticipation



MANUAL SELECTION

Features: 46 R^2: 0.85

Better general fit



5.
Business Conclusions

Optimization Using Data

Maintenance & Repair:

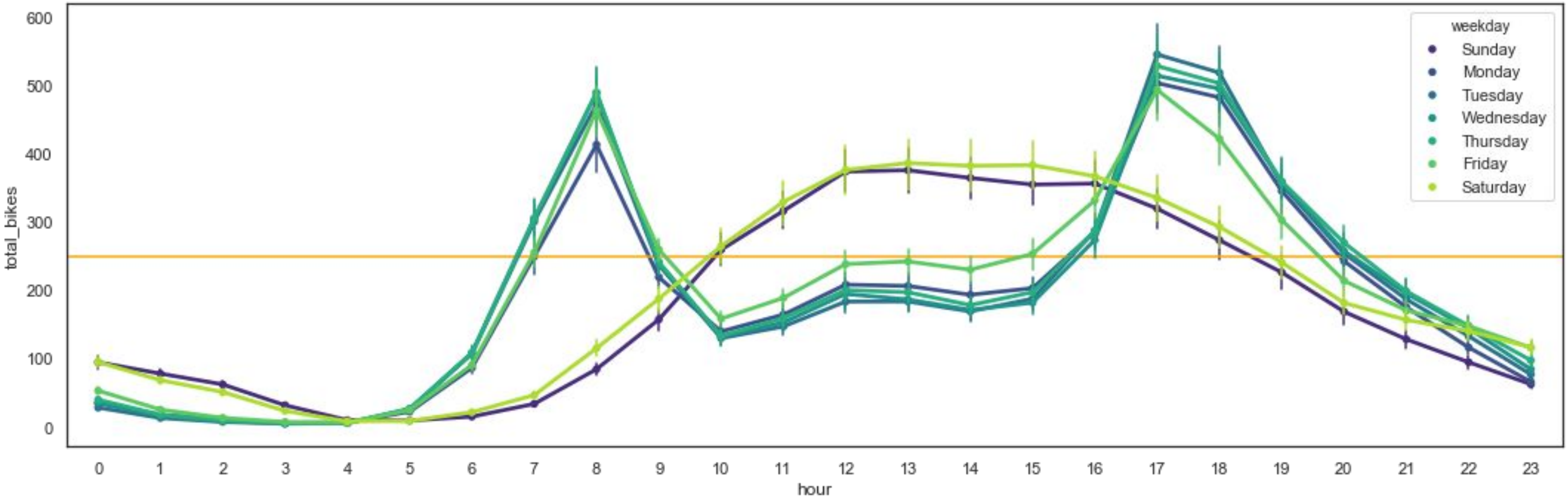
Data driven approach to optimize processes to keep bikes and docks in good repair, safe, and available.

Adapting Technologies for Future Usage:

Optimizing current operations, and the “bike valet service.”

Determining Peak Times

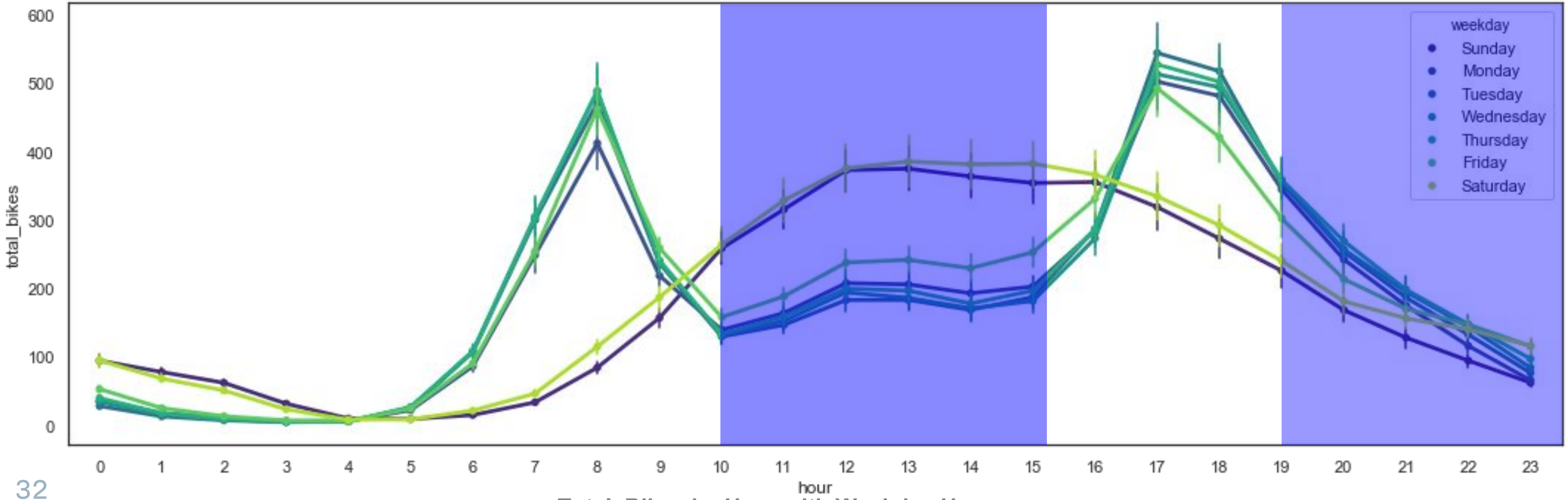
- Peak times based on mean + 31.5%
- Process allows model flexibility
- Additional data will adapt to model



Total_Bikes by Hour with Weekday Hue

Peak Times + Maintenance Weekdays

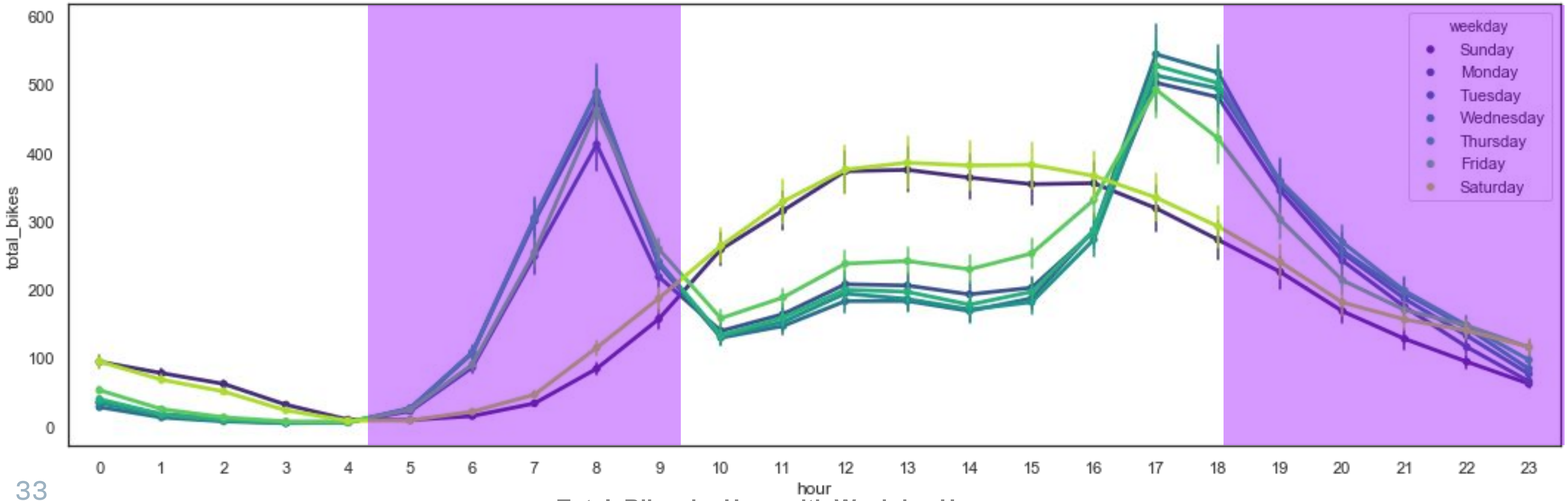
- Weekdays/Commuting-highest usage
- Peak hours for determining maintenance time
- Goal: Least disturbance to business



Total_Bikes by Hour with Weekday Hue

Peak Times + Maintenance Weekends

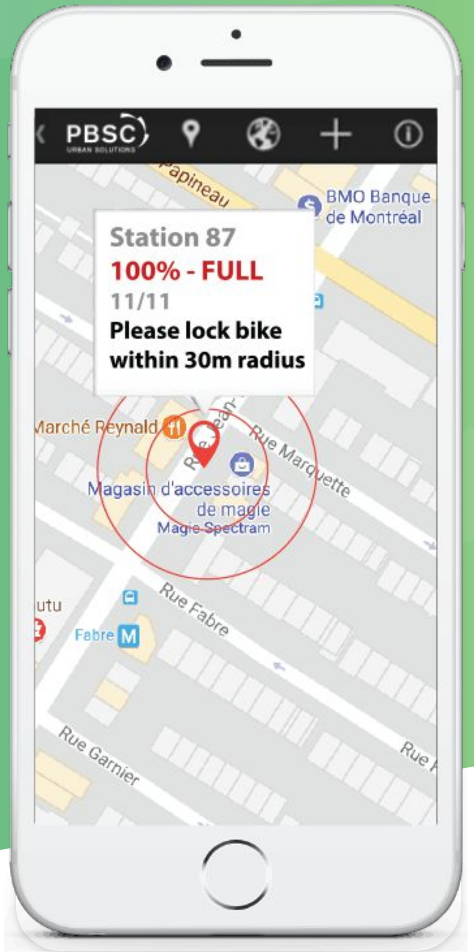
- Weekend-lower usage
- Peak hours different from weekday
- Goal: Least disturbance to business



Total_Bikes by Hour with Weekday Hue

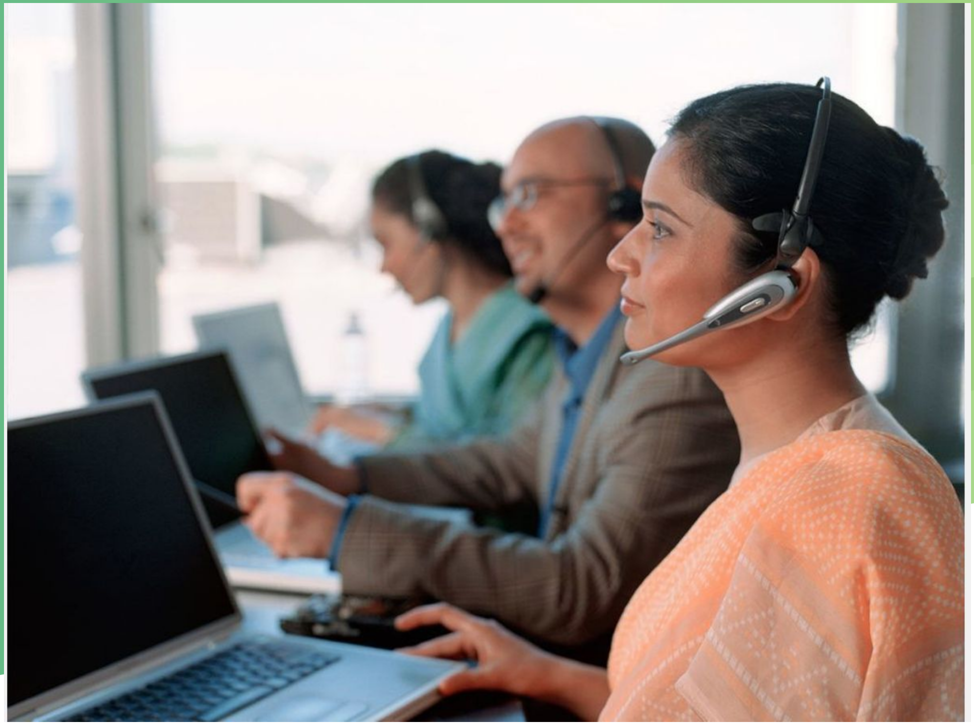
Optimizing Operations

- Rebalancing
- Bike Valet Service
- Geofencing/Station Availability



Peak Times and Growth Optimization

- Use models in conjunction with other departments
- Avg. time increases can provide insight on inventory
- Optimize inventory based on trends



Bike Sharing in Washington D.C.