HONORARY PATRONAGE
RECTOR OF

SZKOŁA GŁÓWNA HANDLOWA W WARSZAWIE

IMPROVING PREDICTIVE MODELING THE WAY FOR PERSONAL CARRIER

ADAM KARWAN, PHD

7 NOVEMBER 2017, WARSAW

Kaggle

The Home of Data Science and Machine Learning

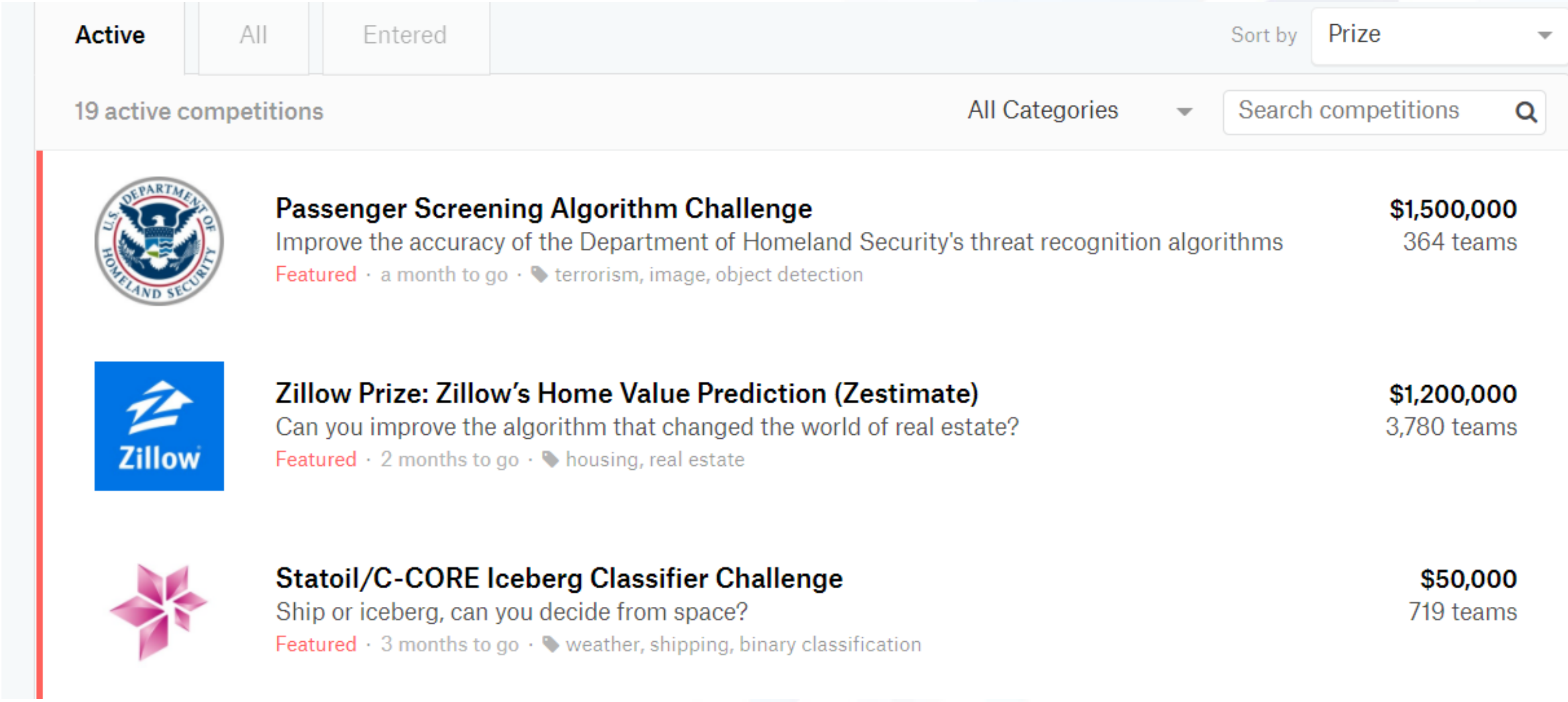
- **WHAT IS KAGGLE** (Data Science as a Competition)

kaggle

What could the world's best analysts find in your data?

- **Data Sets**
- **Contests**

What is Kaggle



- **Self-Learning**
- **Recruitment**
- **Algorithm for Companies**

BOEING

Boeing Global Services
Digital Aviation & Analytics Lab
Gdansk

BIKE SHARING CHALLENGE

https://www.kaggle.com/c/bike-sharing-demand

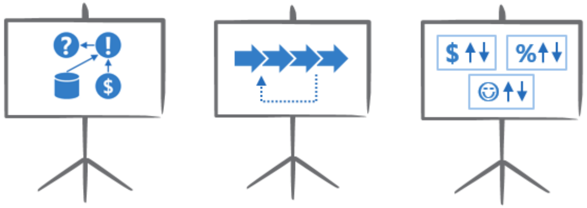
GOAL

combine **historical usage patterns** with **weather data** in order to **forecast bike rental demand** in the Capital Bike Share Program in Washington, D.C.

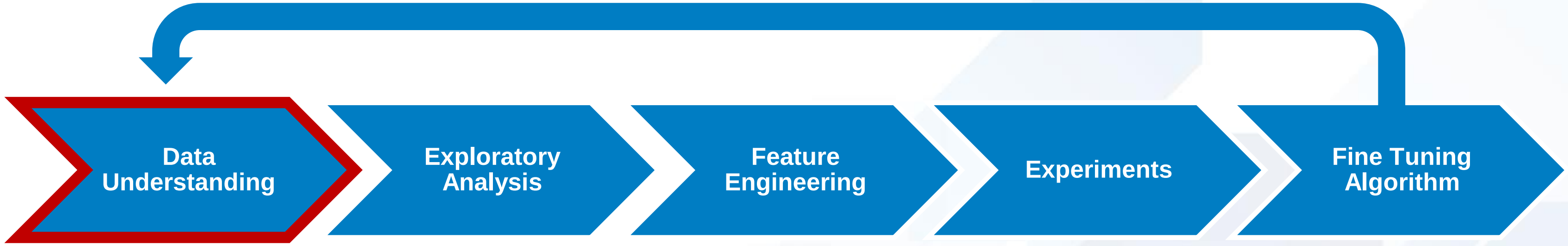
Algorithm + **Presentation**

- **Personal Goal –** create model in **TOP100** best solutions in 7 days

- **How to properly present a Data Mining project?**
  - **Start with big picture**
  - **Overview of process**
  - **Show the main outcome**

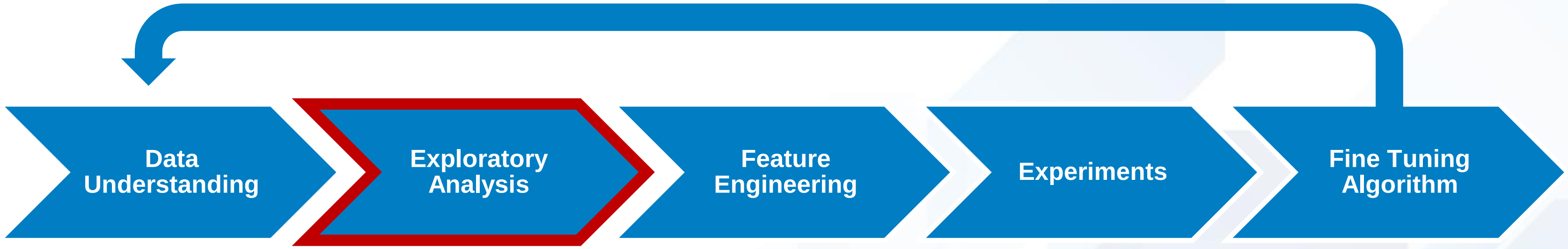


Source http://algolytics.com/data-visualization-essentials-for-data-scientists

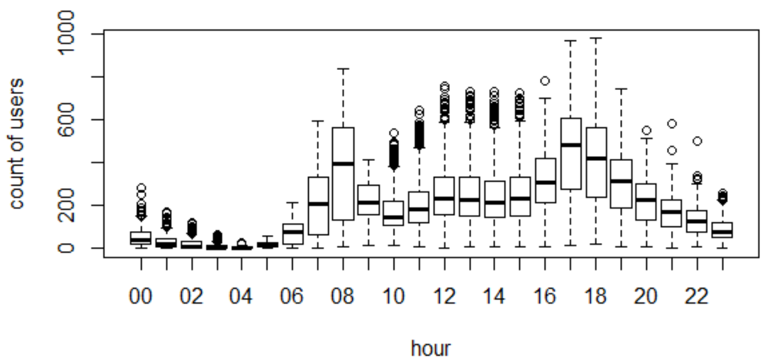


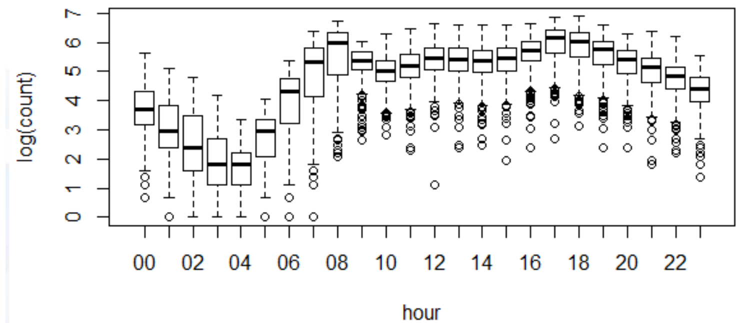
No	Attribute	Description
1	***datetime***	hourly date +timestamp
2	***season***	1 = spring 2 = summer 3 = fall 4 = winter
3	***holiday***	whether the day is considered a holiday (0; 1)
4	***workingday***	whether the day is neither a weekend nor holiday (0; 1)
5	***weather***	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
6	***temp***	temperature in Celsius
7	***atemp***	"feels like" temperature in Celsius
8	***humidity***	relative humidity
9	***windspeed***	wind speed
10	***casual***	number of non-registered user rentals initiated
11	***registered***	number of registered user rentals initiated
12	***count***	number of total rentals

TRAIN	10886
TEST	6493



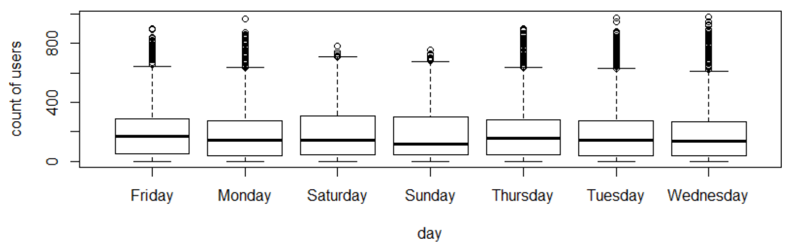
- **Hourly**



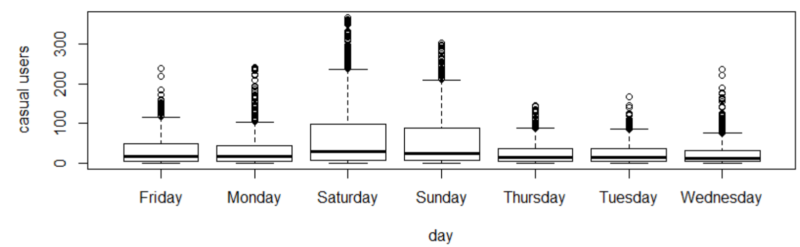


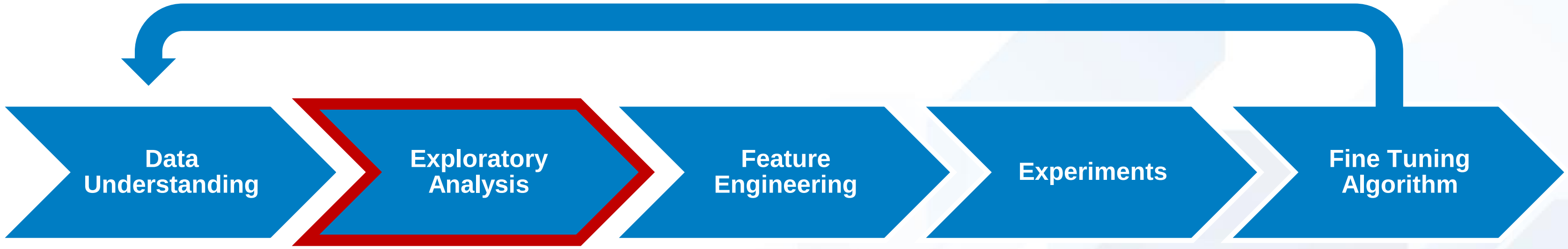
- **Daily**

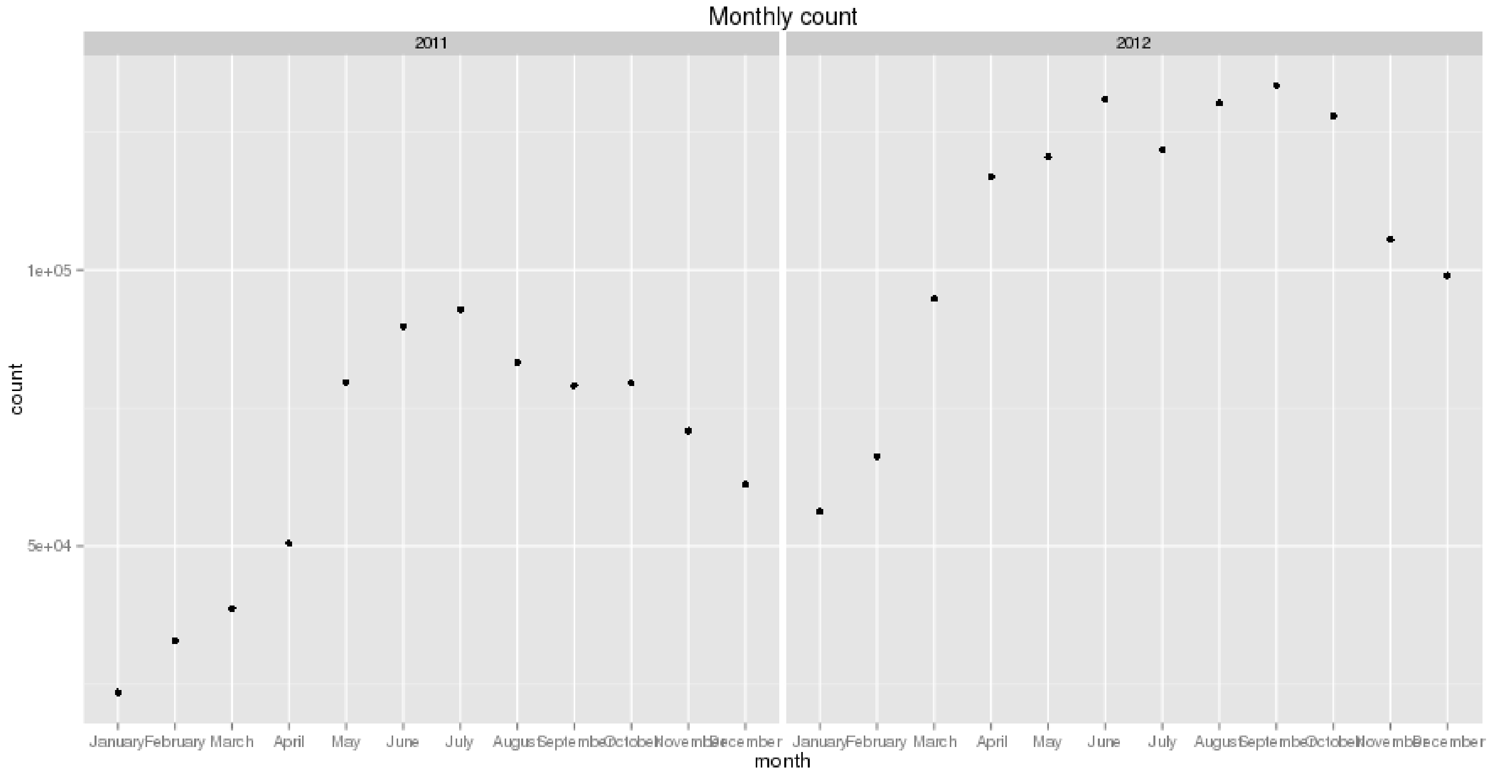
registered

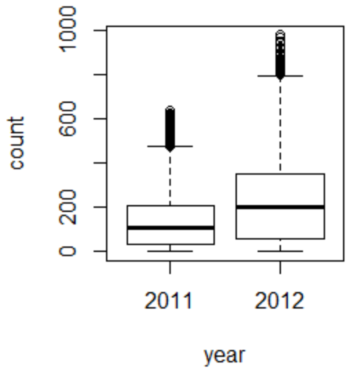


casual



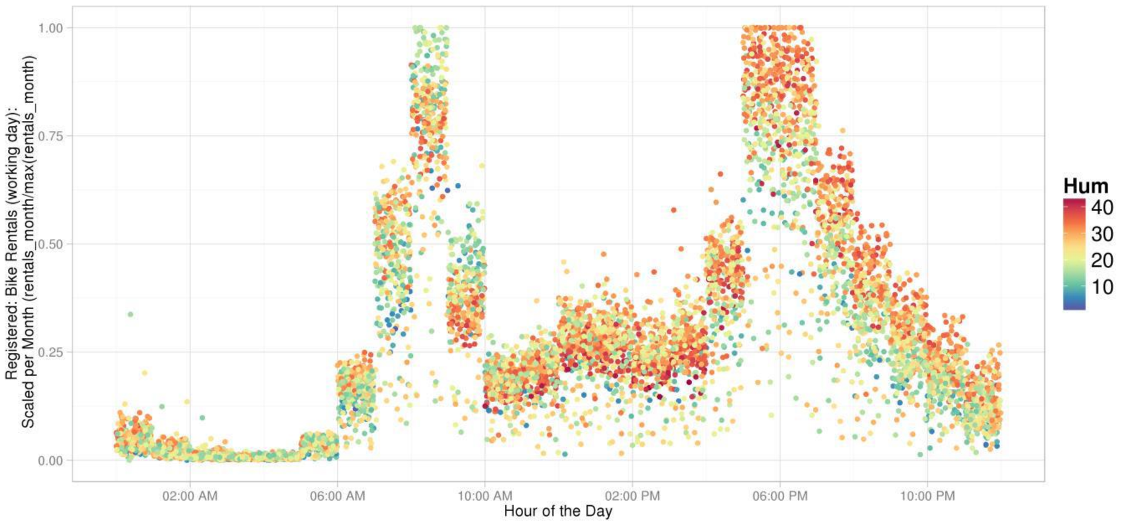

Data Understanding
Exploratory Analysis
Feature Engineering
Experiments
Fine Tuning Algorithm

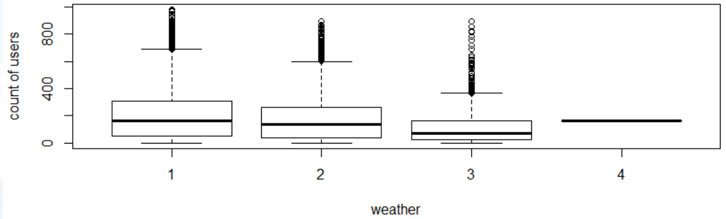

Monthly count
2011
2012
count
1e+05
5e+04
month
January February March April May June July August September October November December

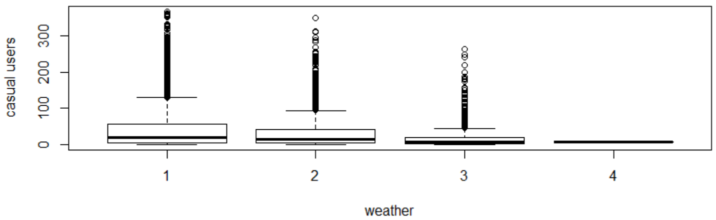

count
0
200
600
1000
2011
2012
year

Weather, Temperature, Humidity

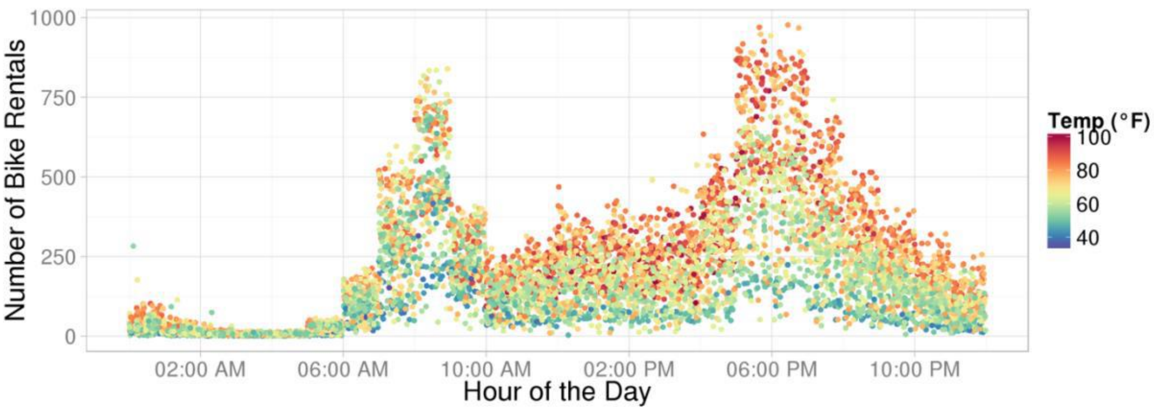
On workingdays, any deducible effect of humidity, by any chance...?
(taking the bike to work no matter what, but dry if back home...?)

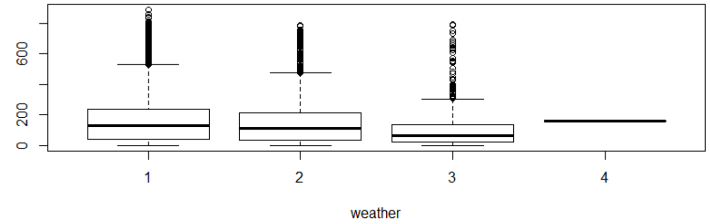






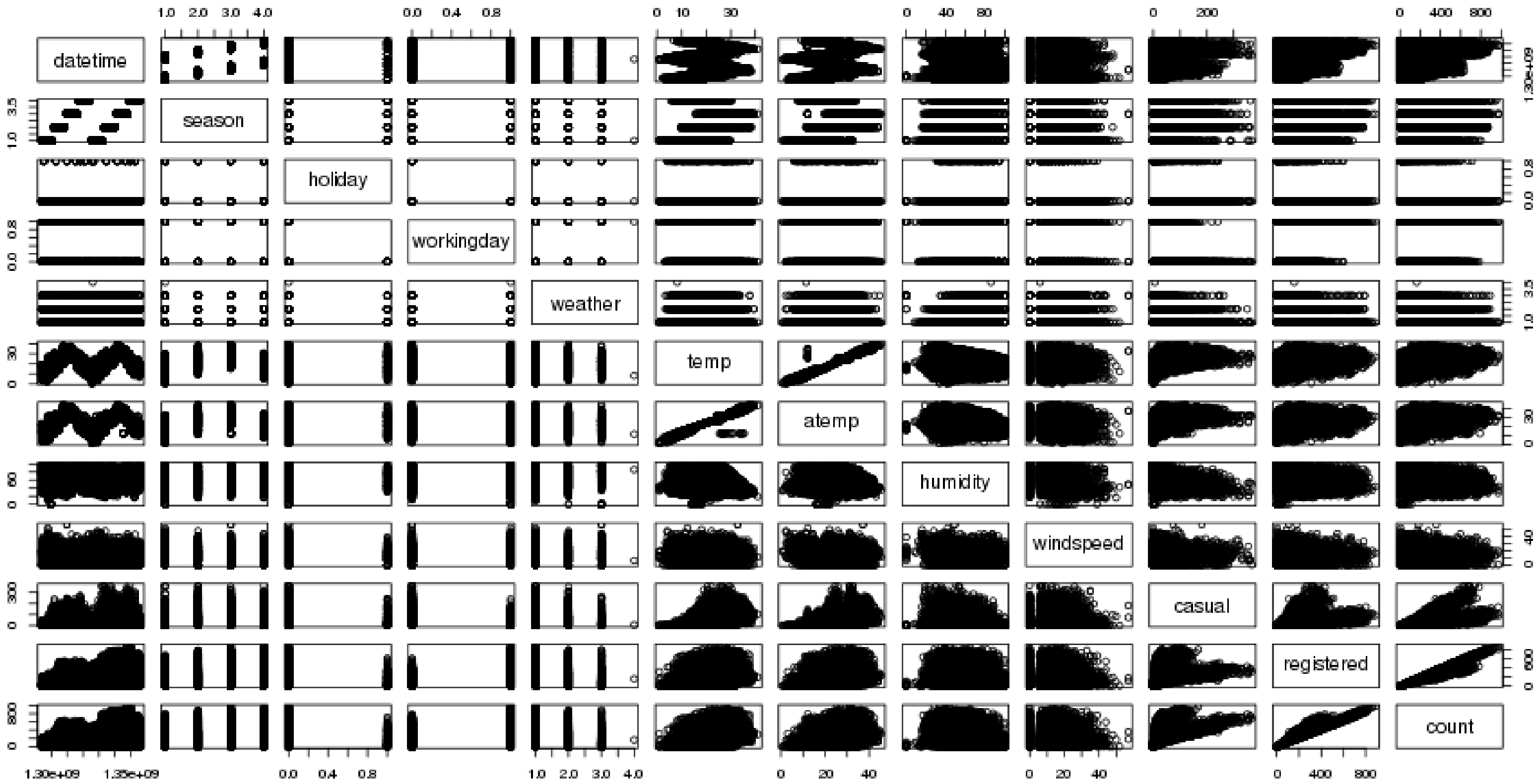
On workdays, most bikes are rented on warm mornings and evenings

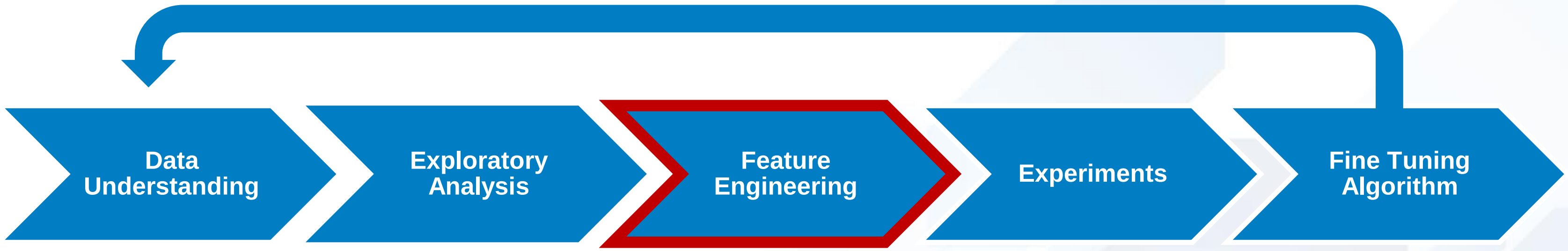




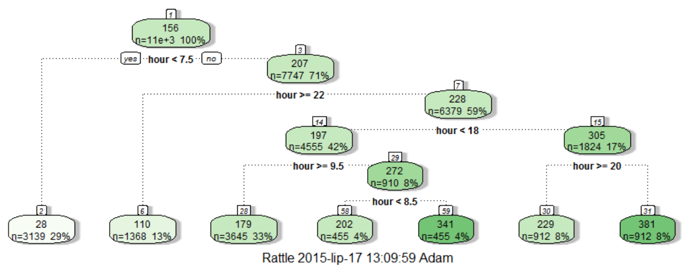
Correlations

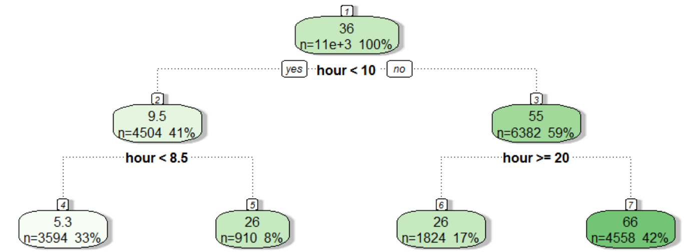
	train.registered	train.casual	train.count	train.temp	train.humidity	train.atemp	train.windspeed
train.registered	1.00	0.50	0.97	0.32	-0.27	0.31	0.09
train.casual	0.50	1.00	0.69	0.47	-0.35	0.46	0.09
train.count	0.97	0.69	1.00	0.39	-0.32	0.39	0.10
train.temp	0.32	0.47	0.39	1.00	-0.06	0.98	-0.02
train.humidity	-0.27	-0.35	-0.32	-0.06	1.00	-0.04	-0.32
train.atemp	0.31	0.46	0.39	0.98	-0.04	1.00	-0.06
train.windspeed	0.09	0.09	0.10	-0.02	-0.32	-0.06	1.00

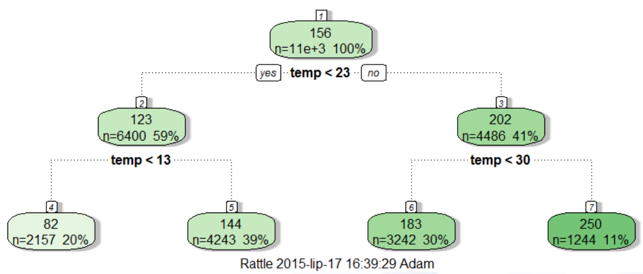


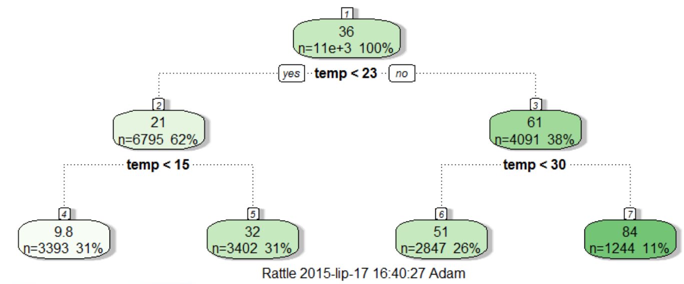


- *dp_reg*
- *dp_cas*
- *temp_reg*
- *temp_cas*

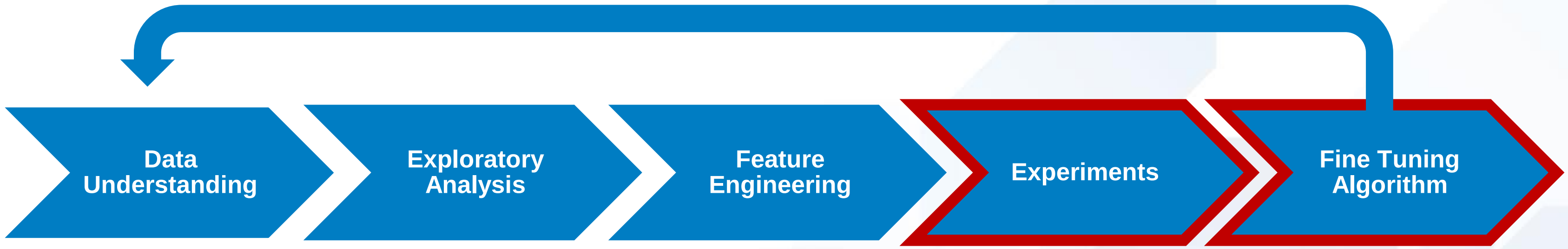








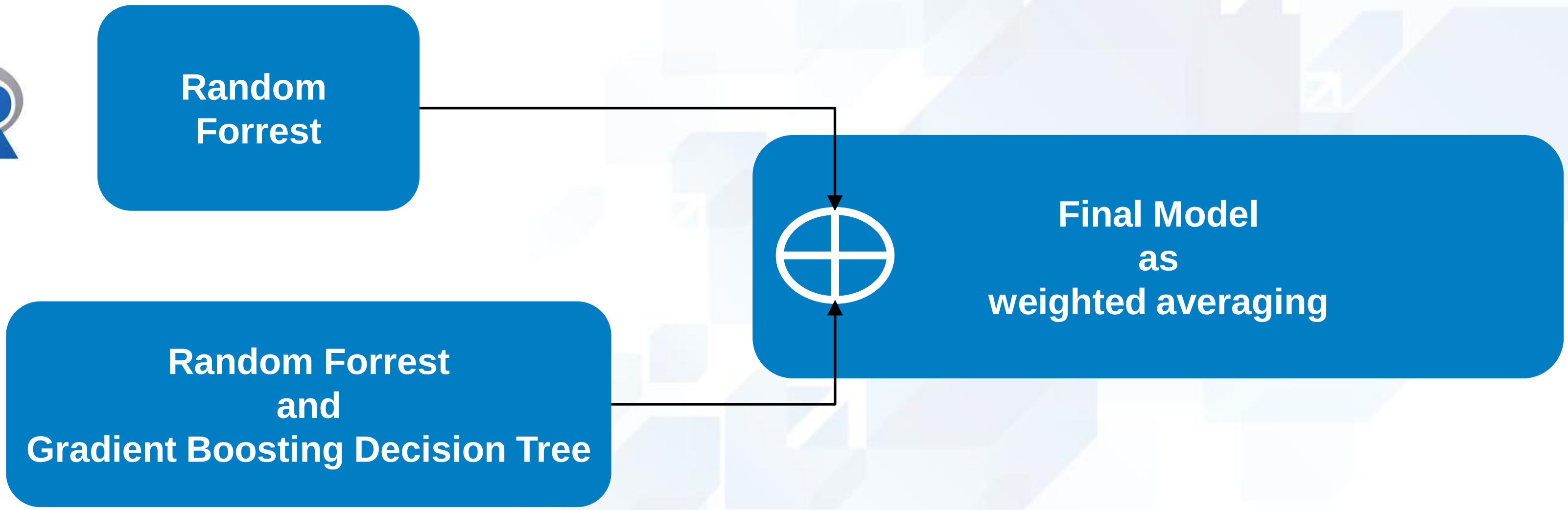
- *day_type*
  - Holiday [holiday=0 and workingday=0]
  - Weekend [holiday=1]
  - Working Day [holiday=0 and workingday=1]
- *year_part*
  - (1 – 8) from first quarter of 2011 till fourth of 2012



- **R model** bases on two separately computed **Random Forests** for *casual* and *registered* users. Final result is a sum of predicted values daily. Due to differences in *count* of users *hourly* we use logarithm to normalize values. Moreover each model was trained on 250 trees. In first model we estimate number of registered users therefore *dp_reg* and *temp_reg* are used, analogously *dp_cas* and *temp_cas* for second one. Other attributes used: *hour, day, day_type, holiday, season, year, year_part, weekend, workingday, atemp, humidity, weather, windspeed.*

- **Python model** bases on a combination of ***RF (Random Forrest) and GBDT (Gradient Boosting Decision Trees).*** Twelve attributes are used: *hour, day, holiday, season, weekday, workingday, year, atemp, temp, humidity, weather, windspeed. Year* is normalized by substraction 2011. GBDT is computed on 100 and RF on 1000 trees. Before computing final result we compute average regression of two instances for each approach RF and GBDT. For estimated variables logarithm is used to normalize results.

Ensamble Method



Final result **0.3704** that was **64** place out of **3252**

63	↓11	Dhruv Singal	0.37011	6	Sat, 11 Apr 2015 23:34:25 (-9.1d)
-		**a.karwan**	**0.37040**	-	**Tue, 21 Jul 2015 20:24:47** Post-Deadline
Post-Deadline Entry If you would have submitted this entry during the competition, you would have been around here on the leaderboard.					
64	↓11	prashkr	0.37041	6	Thu, 09 Apr 2015 20:53:07 (-23.1h)

JEPPESEN

A BOEING COMPANY

BOEING

Boeing Global Services

Digital Aviation & Analytics Lab

Gdansk

DATA SCIENCE in AVIATION INDUSTRY

- **PREDICTIVE MX**
- **POST FLIGHT ANALYTICS**
- **REAL TIME MACHINE LEARNING**

Titanic Machine Learning from Disaster

https://www.kaggle.com/c/titanic

More likely to survive

- Females
- Children
- 1st Class Passengers
- Traveling with Family

More likely to perish

- Males
- Adults
- 2nd and 3rd Class Passengers
- Traveling alone

Trevor Stephens

Regular Data Scientist,
Occasional Blogger.

San Francisco, CA

Website

Twitter

LinkedIn

Github

43	↑499	**Total Recall**	0.81340	8

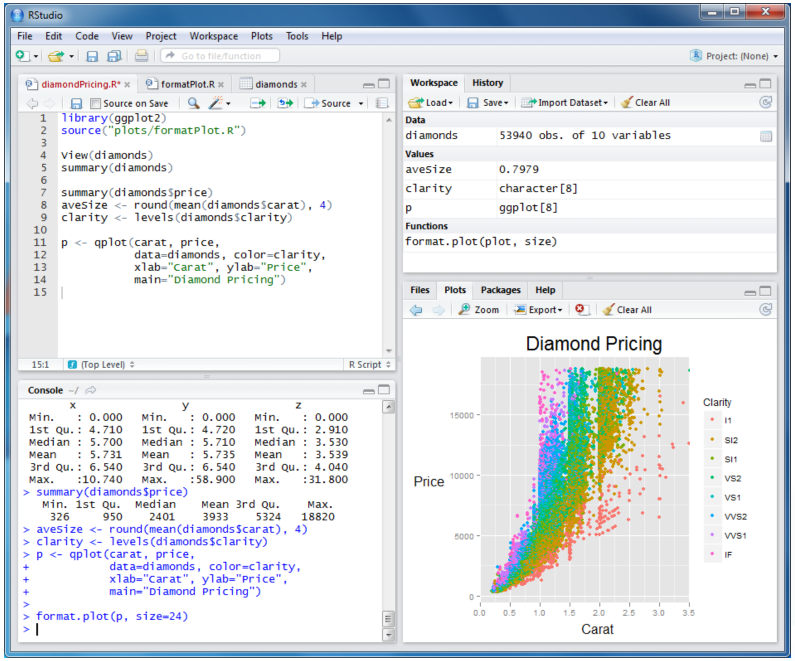
Your Best Entry
You improved on your best score by 0.01914.
You just moved up 219 positions on the leaderboard.

Solution in Top 5%

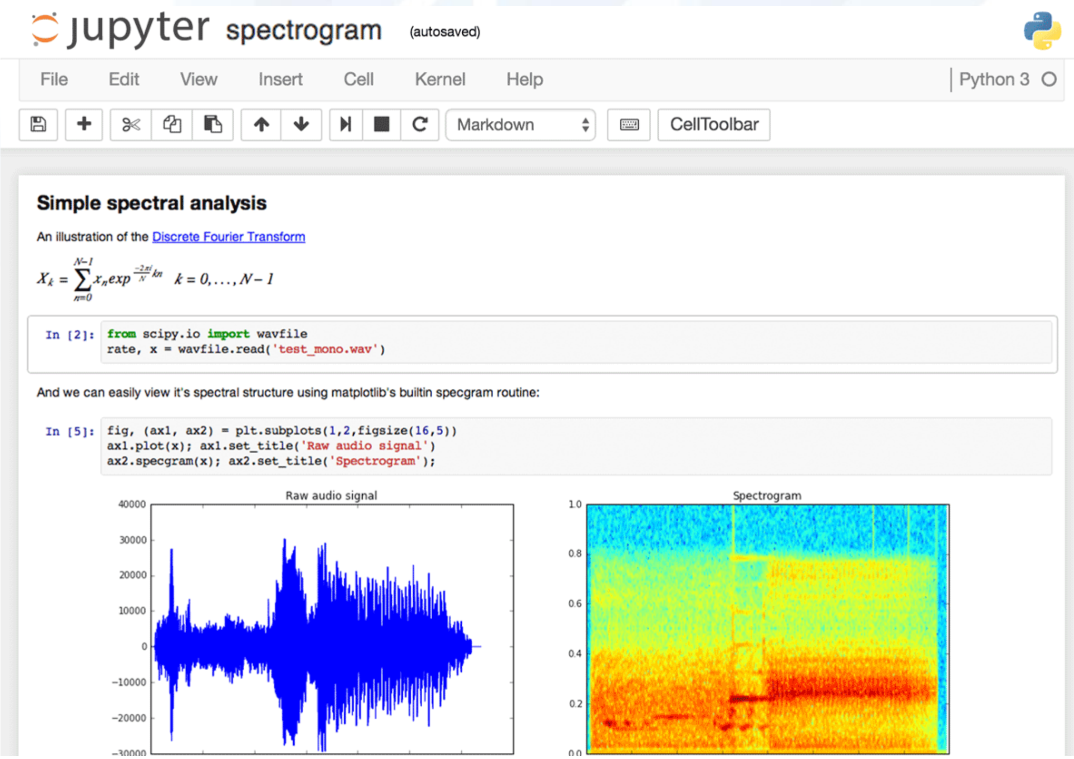
Tutorial

http://trevorstephens.com/kaggle-titanic-tutorial/getting-started-with-r

R Studio, Python Anaconda, Jupyter



R Studio



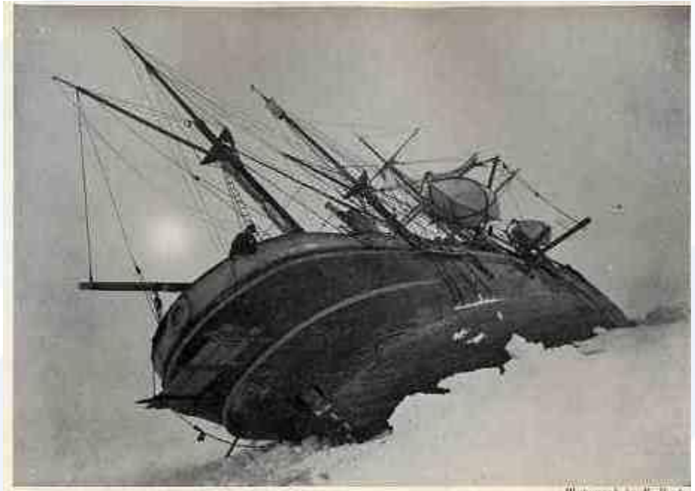
jupyter

spyder

ANACONDA

Kaggle

Tips and tricks



Photograph by F. Hurley
WITHIN A FEW SECONDS SHE HEELED OVER UNTIL SHE HAD A LIST OF THIRTY DEGREES TO PORT

- Be patient
- Understand data
- Think more, code less
- Follow other approaches (Kaggle Forum, Github, Slideshare)
- Fit appropriate algorithm to data

355	↓31	Luminous Logic	0.41089	16	Sat, 27 Dec 2014 18:37:49 (-24.7h)
-		**a.karwan**	**0.41091**	-	**Mon, 20 Jul 2015 14:25:45** Post-Deadline
Post-Deadline Entry If you would have submitted this entry during the competition, you would have been around here on the leaderboard.					
356	↓31	rachel wang	0.41092	15	Sun, 12 Apr 2015 01:57:48

1597	new	Chai Kothari	0.49523	19	Fri, 29 May 2015 03:36:57 (-0.5h)
-		**a.karwan**	**0.49523**	-	**Mon, 20 Jul 2015 11:18:56** Post-Deadline
Post-Deadline Entry If you would have submitted this entry during the competition, you would have been around here on the leaderboard.					
1598	↓103	batussi	0.49523	2	Mon, 24 Nov 2014 14:29:33 (-0.1h)