Photo Source: https://www.bizjournals.com/washington/blog/2014/07

Capital Bikeshare Strategy

- A Support Vector Regression Approach

Presenter: Mingyuhui Liu (Jane)

Principle Instructor: Dr. Benjamin Harvey

Agenda

- Introduction
- Data Preprocessing:
  - Descriptive data and figures
  - Outliers Detection
- Data Analysis
- Business Strategy Suggestions

Introduction

- Data:
  - Capital Bikeshare rental counts.
  - Other information integrated.

Photo sources: Ben Schumin (https://en.wikipedia.org/wiki/Capital_Bikeshare)

- Objective:
  - Find the correlation of rental counts and explanatory variables
  - What can be done to increase the rental counts
  - Audience

- Methodology:
  - Support Vector Regression (SVR)
  - Parameter Tuning

Photo sources: David Alpert (https://ggwash.org/view)

Preprocessing: Descriptive Statistics and Figures

- Original Data: 17 columns, 17,379 instances

- instant: record index

- dteday : date

- season : season (1:spring, 2:summer, 3:fall, 4:winter)

- yr : year (0: 2011, 1:2012)

- mnth : month (1 to 12) ⟶ Recode "month": 1 ~ 24

- hr : hour (0 to 23)

- holiday : weather day is holiday or not (extracted from [Web Link])

- weekday : day of the week

- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.

Preprocessing: Descriptive Statistics and Figures

- Original Data:

- weathersit :
  - 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
- temp : Normalized temperature in Celsius.
  - The values are derived via **(t-t_min)/(t_max-t_min),** t_min=-8, t_max=+39 (only in hourly scale)
- atemp: Normalized feeling temperature in Celsius.
  - The values are derived via **(t-t_min)/(t_max-t_min),** t_min=-16, t_max=+50 (only in hourly scale)

Preprocessing: Descriptive Statistics and Figures

- Original Data:

- hum: Normalized humidity. The values are divided to 100 (max)

- windspeed: Normalized wind speed. The values are divided to 67 (max)

- casual: count of casual users

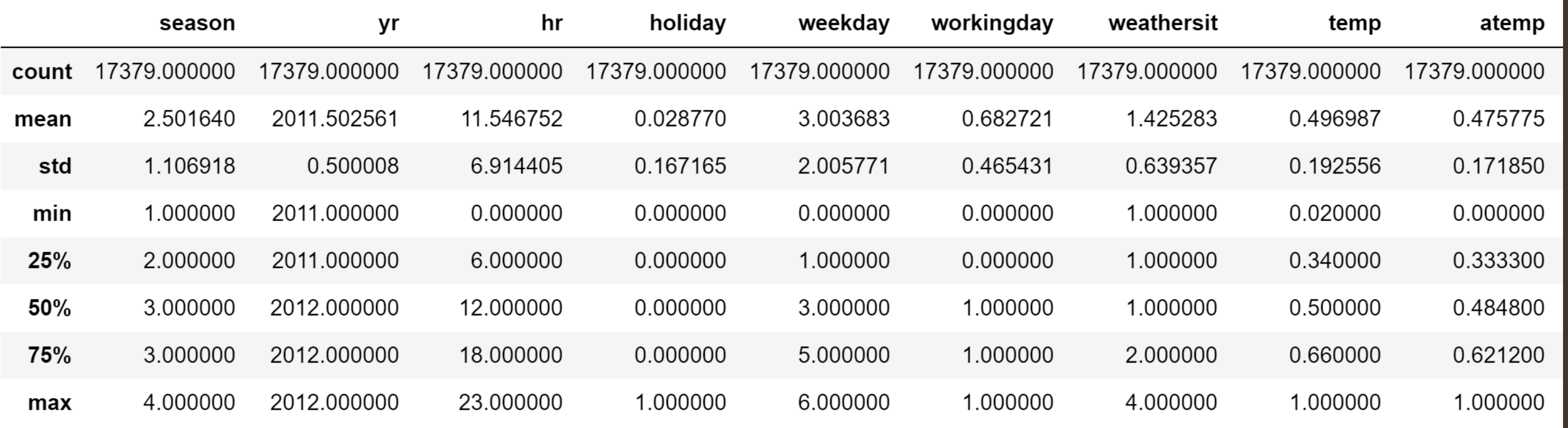
- registered: count of registered users

- cnt: count of total rental bikes including both casual and registered

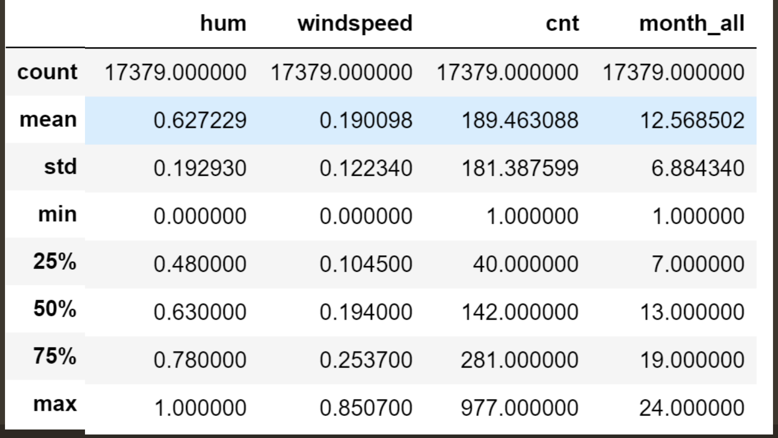
Target

Preprocessing: Descriptive Statistics and Figures

- Examples from the Cleaned Data:



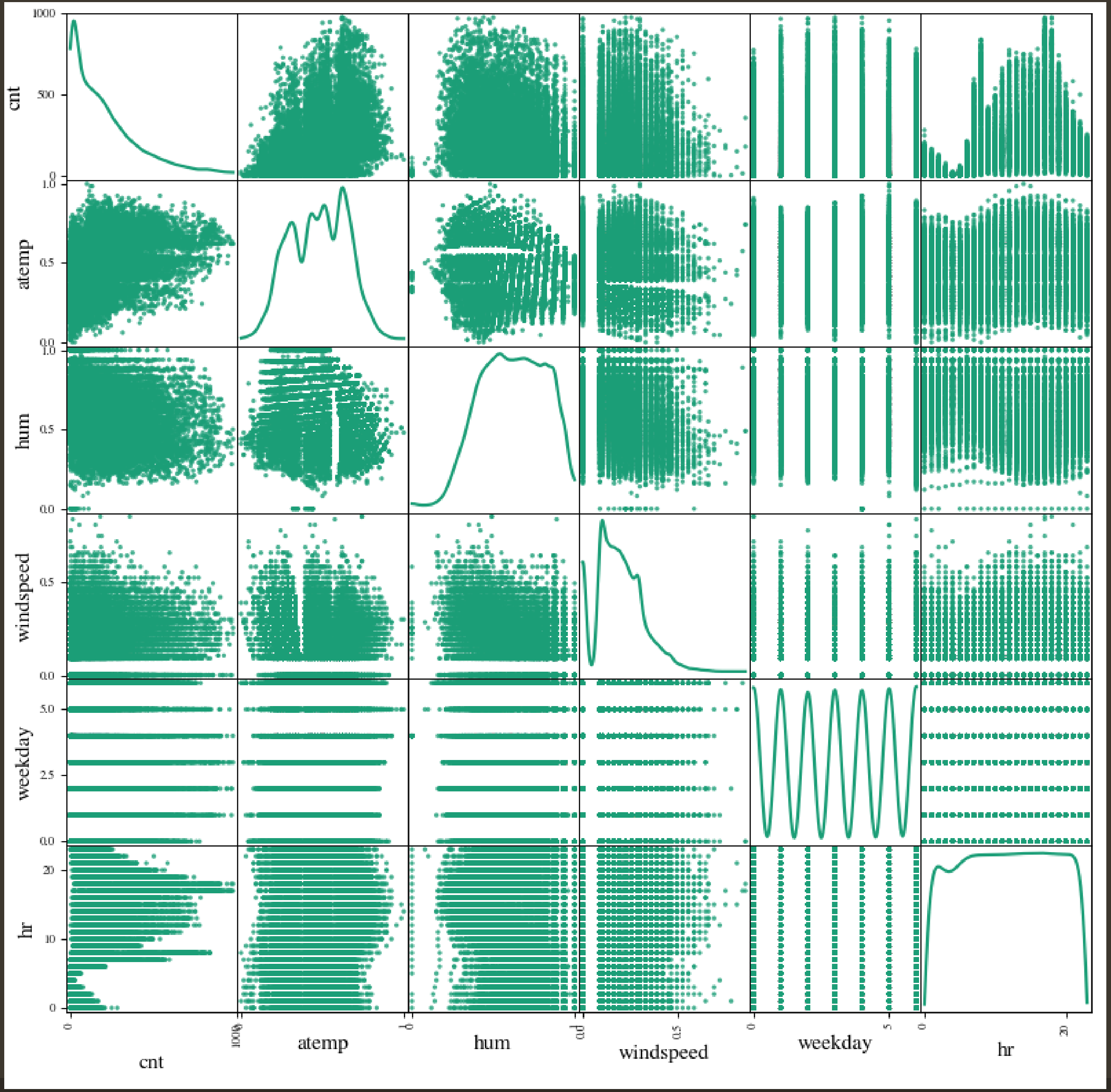
	season	yr	hr	holiday	weekday	workingday	weathersit	temp	atemp
count	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	2.501640	2011.502561	11.546752	0.028770	3.003683	0.682721	1.425283	0.496987	0.475775
std	1.106918	0.500008	6.914405	0.167165	2.005771	0.465431	0.639357	0.192556	0.171850
min	1.000000	2011.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.020000	0.000000
25%	2.000000	2011.000000	6.000000	0.000000	1.000000	0.000000	1.000000	0.340000	0.333300
50%	3.000000	2012.000000	12.000000	0.000000	3.000000	1.000000	1.000000	0.500000	0.484800
75%	3.000000	2012.000000	18.000000	0.000000	5.000000	1.000000	2.000000	0.660000	0.621200
max	4.000000	2012.000000	23.000000	1.000000	6.000000	1.000000	4.000000	1.000000	1.000000



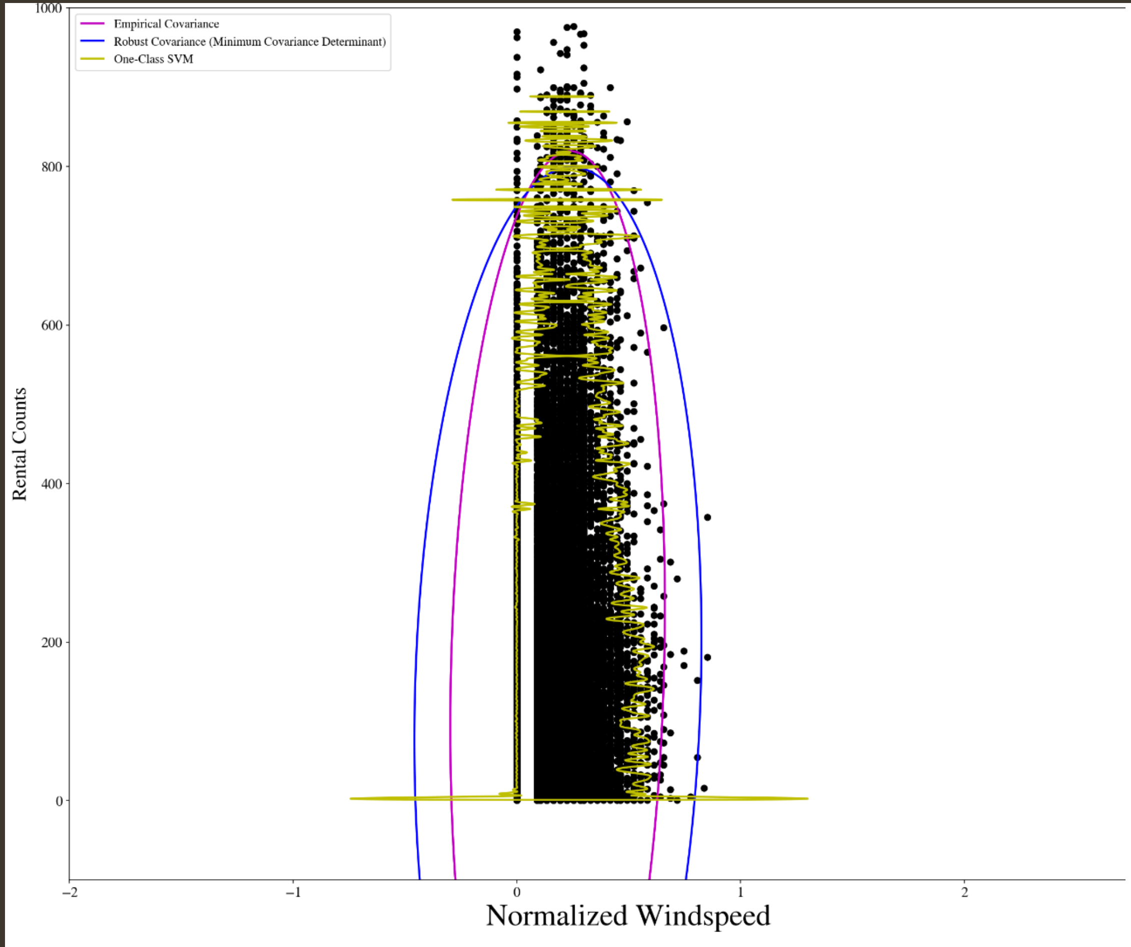
	hum	windspeed	cnt	month_all
count	17379.000000	17379.000000	17379.000000	17379.000000
mean	0.627229	0.190098	189.463088	12.568502
std	0.192930	0.122340	181.387599	6.884340
min	0.000000	0.000000	1.000000	1.000000
25%	0.480000	0.104500	40.000000	7.000000
50%	0.630000	0.194000	142.000000	13.000000
75%	0.780000	0.253700	281.000000	19.000000
max	1.000000	0.850700	977.000000	24.000000

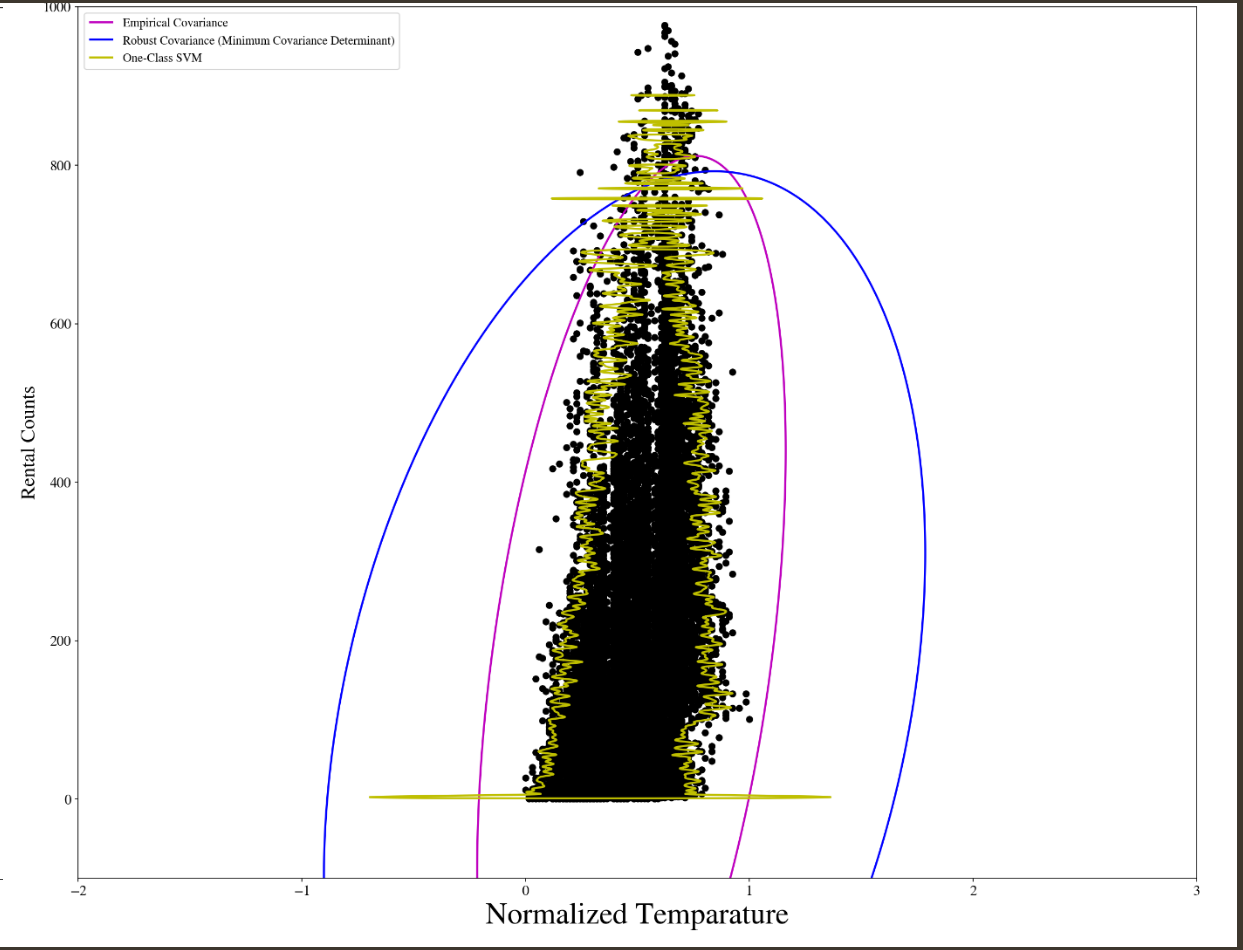
Preprocessing: Descriptive Figures

- Matrix
  - Treat as continuous
  - No clear correlation
  - SVR Approach



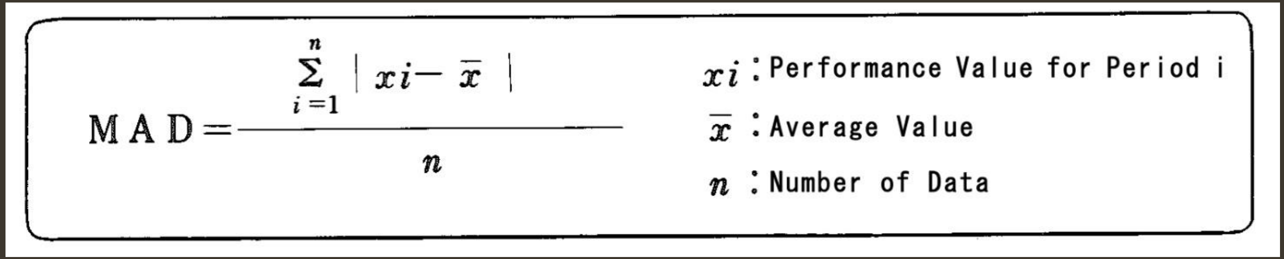
Preprocessing: Outlier Detection





Outlier Detection

- Median Absolute Deviation (MAD):



$$\mathrm{MAD} = \frac{\sum_{i=1}^{n} \left| xi - \bar{x} \right|}{n}$$

xi : Performance Value for Period i

$\bar{x}$: Average Value

n : Number of Data

- On the dependent variable – “count”:
  - Detected 267 outliers
  - Drop the outliers
  - As a comparison with the whole dataset

Analysis: SVR - fitting

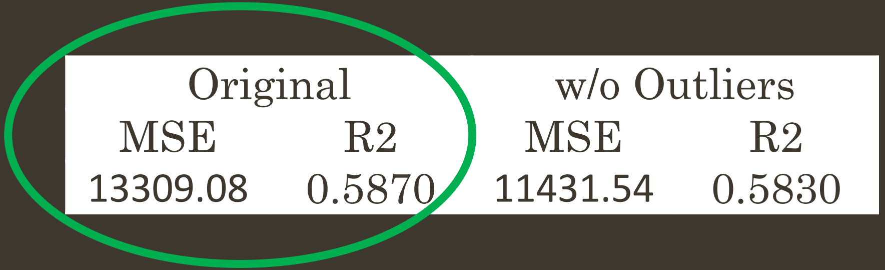
- Model Building:
  - Outliers:
    - w/ Vs w/o
  - Variables Selection:
    - **Main:** "atemp", "weathersit", "workingday", "hr"
    - **Random**: "season", "month_all", "windspeed“
  - Training Vs Testing:
    - 0.12 to 0.33
  - Kernel:
    - “rbf”, “linear”

Model	Ratio	Gaussian		Linear	
		MSE	R2	MSE	R2
atemp', 'weathersit', 'workingday', 'hr', 'season', 'month_all', 'windspeed'	33.33%	14561.58	0.5478	22953.20	0.2872
	25.00%	14009.37	0.5676	23207.64	0.2837
	20.00%	13496.47	0.5770	22915.55	0.2818
	16.67%	13357.53	0.5810	22976.38	0.2792
	14.29%	13326.43	0.5846	23303.85	0.2736
	12.00%	13309.08	0.5870	23140.17	0.2749

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y_i} - Y_i)^2$$

Analysis: SVR – parameter tuning (1)

- Model Selected: 7 explanatory variables
  - 'atemp', 'weathersit', 'workingday', 'hr', 'season', 'month_all', 'windspeed’
  - Testing ratio = 0.12
  - With Outliers



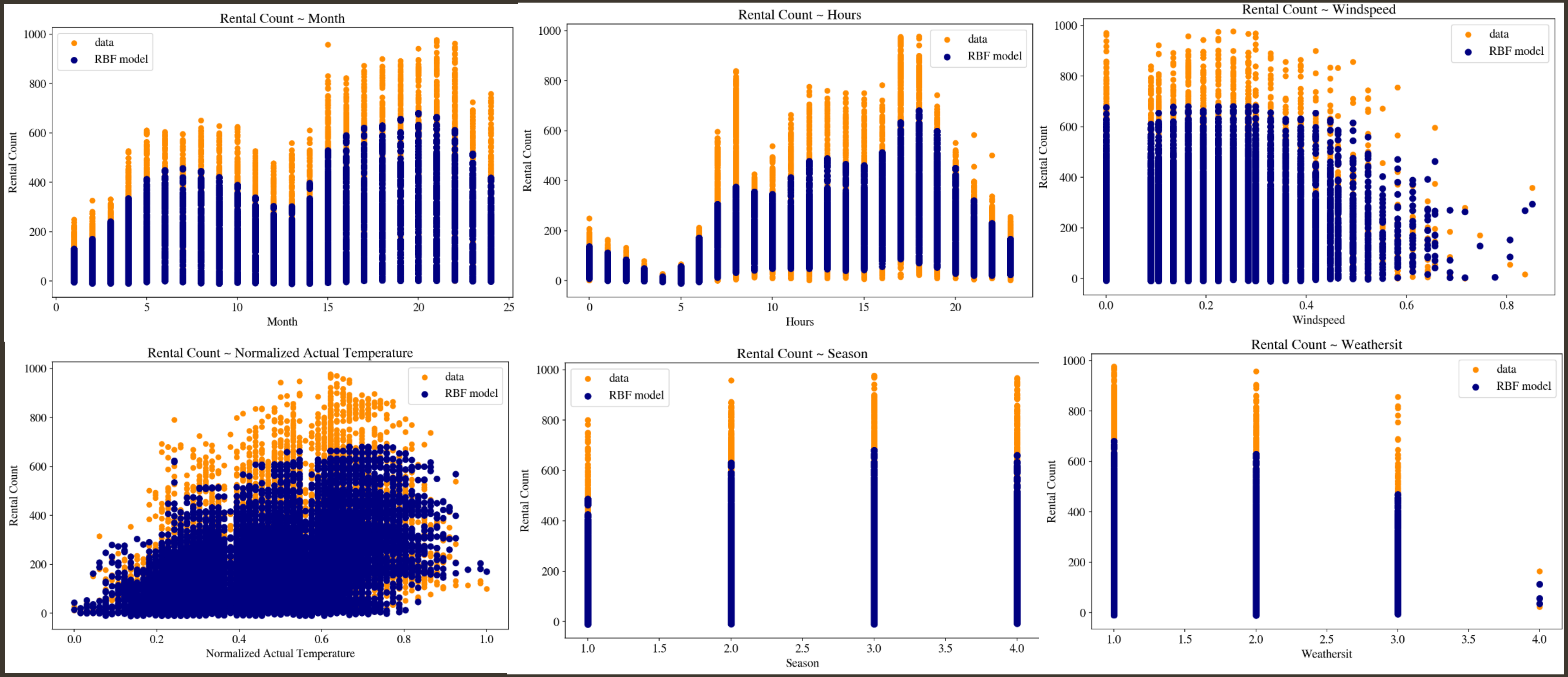
Original		w/o Outliers	
MSE	R2	MSE	R2
13309.08	0.5870	11431.54	0.5830

- Parameter Tuning:
  - parameters = {'kernel':['rbf'], 'C':[1, 10], 'gamma': [0.14, 0.1]}
  - GridSearchCV (svr, parameters)
  - **Results**:
    - SVR(C=10, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma=0.1, kernel='rbf', max_iter=-1, shrinking=True, tol=0.001, verbose=False)

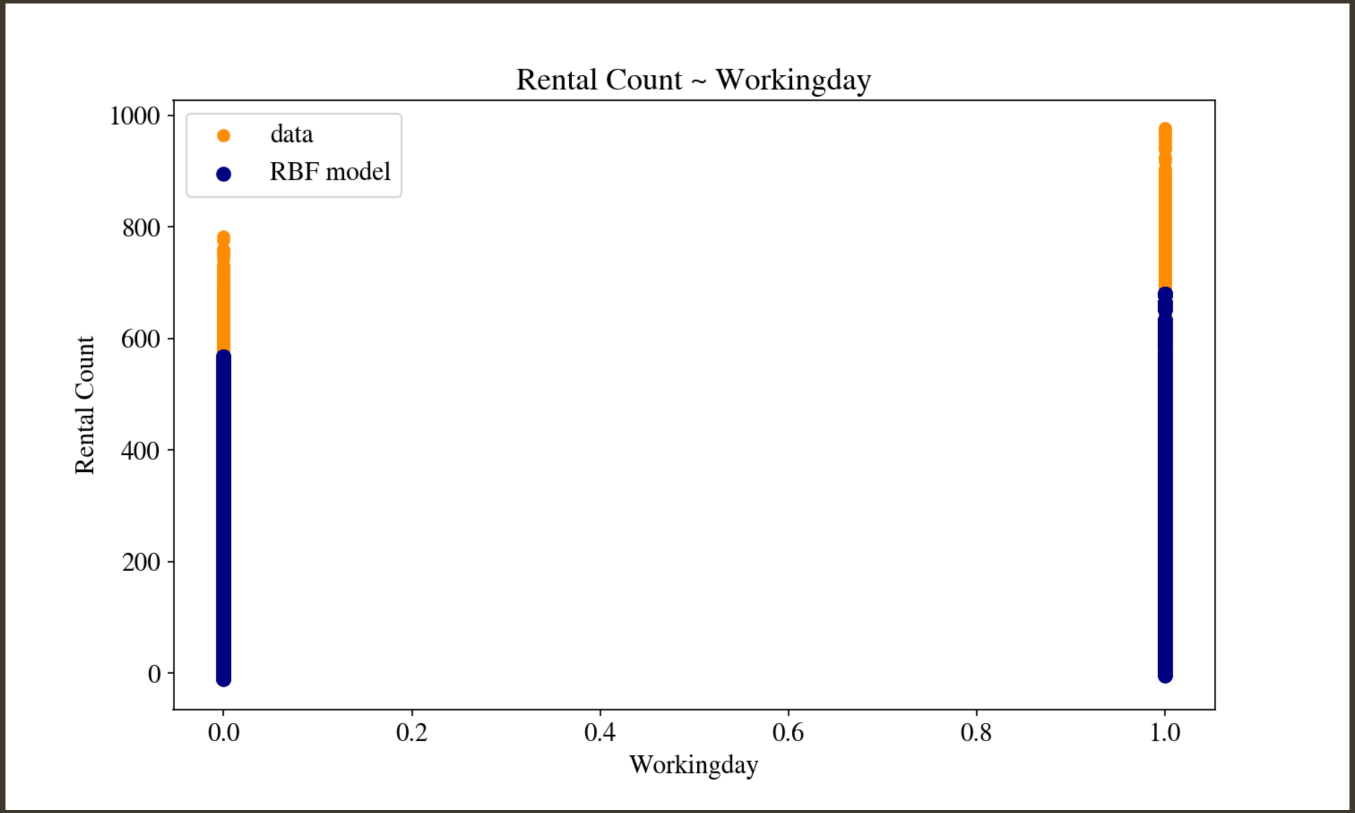
Analysis: SVR – parameter tuning (2)

- GridSearch Best Parameter Results:
  - MSE: 6646.88 (Improved: 50.06%)
  - R2: 0.7937 (Improved: 35.28%)

Analysis: SVR – result visualization



Analysis: SVR – result visualization



Analysis: SVR – ANOVA on “Workingday”

- One-way ANOVA:
  - Null Hypothesis: mean(non-workingday) = mean(workingday)
  - P-value: 6.52E-05
    - Significant
    - Reject the Null
- More ANOVA:
  - Grouped by workingday AND season: p-value = 2.65E-10

This is my focus on the Business Strategy.

Business Strategy: Information Extraction

- General Idea from Business Perspective:
  - Assumption: Tourist Vs Commuter
  - Elasticity

- Elasticity Results:
  - Over Actual Temperature, and weather:

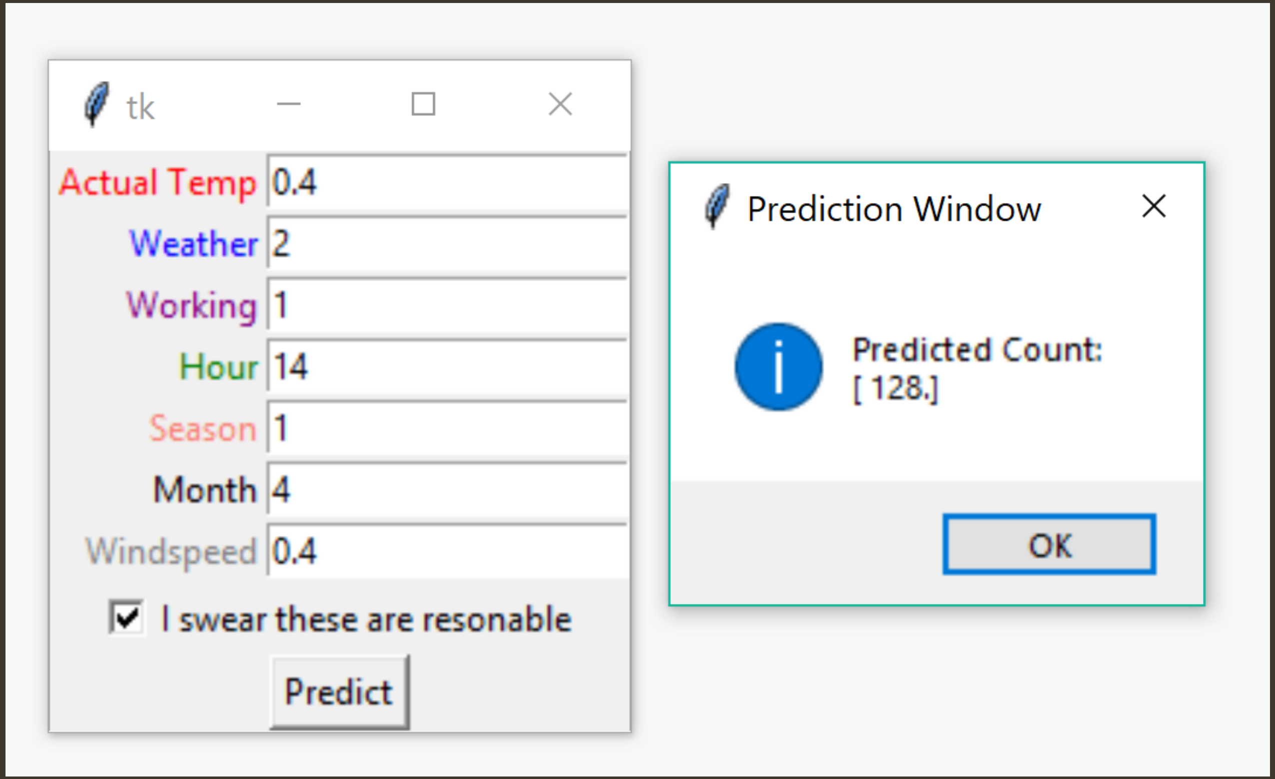
	atemp	% less elasticity	weathersit	% less elasticity
Commuter	0.28732	13.18%	-22.2223	40.43%
Traveler	0.330947		-37.3074	

How a Non-working day look like in DC.

Business Strategy: Suggestion

- Strategy:
  - to utilize the inelasticity of commuters to gain more profit;
  - to motivate the tourist to rent bikes regardless of harsh situations.
- Examples:
  - promote more powerful membership programs to attract more commuters to join the CB services (they are probably going to be solid!);
  - locate the docks nearer to business buildings to boost the number of users;
  - create bonus for registered every-day users;
  - locate more docks near tourist places to target specifically at the visitors;
  - build raincoat vending machine in the docks

One More Application



Actual Temp 0.4

Weather 2

Working 1

Hour 14

Season 1

Month 4

Windspeed 0.4

☑ I swear these are resonable

Predicted Count:
[128.]