Bike Sharing Prediction

JCDS12 - FINAL PROJECT

TEAM SCIKIT-LEARN

Ibnu Yusuf Prakoso

Kelvin Lois

Kevin Octavio Untoro

The Scikit-Learn Team

Ibnu Yusuf Prakoso

iyp36@hotmail.com

https://github.com/iyp36

https://www.linkedin.com/in/ibnu-yusuf-prakoso-37045412b/

Kelvin Lois

kelvinnlois@gmail.com

https://github.com/kelvinnlois

https://www.linkedin.com/in/kelvin-lois-187b06201/

https://math.stackexchange.com/users/322139/si-kucing

Kevin Octavio Untoro

kevinoctavio.u@gmail.com

https://github.com/ricel8168

https://www.linkedin.com/in/kevinou-8168

BACKGROUND

Our Purpose for this Project is to predict Bike Sharing Demand and Interest of users in Washington D.C.
Using the system, it is easier for users and members to rent bicycles. In 2017 there are more than 500 bike-sharing systems around the world with a total of 500 thousand bikes.
Currently, the bike-sharing system plays an important role in traffic, environmental and health issues, therefore there is an increase in customer interest.

The dataset used is from the Capital Bike Share Company Data in the U.S. which contains Data for 2011 and 2012, includes with seasons and weather

Problems

- When is the right time to do Maintenance ?
- How to reduce loss of potential customers?
- What does the company need to do when there's "SPIKE" in Bike Sharing Demand ?

Goals

- Reduce Potential Customer Loss
- Maintenance Schedule
- Predicting Bike Sharing Demand

Data Understanding

Data Description

Definition of each Variables

Pre-processing

- Rescaling
- Add new columns
- Checking Missing Value

Exploratory Data Analysis

- Insights
- Data Visualization

- Is there any missing values?
- Is there any Outliers ? How to deal with it ?
- Do we need to bin/rescale the data?
- What is the target used for Machine Learning?
- What is our metric perfomance for Machine Learning ?

Dataset

instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	0	6	0	1	0.24	0.2879	0.81	0.0000	3	13	16
2	2011-01-01	1	0	1	1	0	6	0	1	0.22	0.2727	0.80	0.0000	8	32	40
3	2011-01-01	1	0	1	2	0	6	0	1	0.22	0.2727	0.80	0.0000	5	27	32
4	2011-01-01	1	0	1	3	0	6	0	1	0.24	0.2879	0.75	0.0000	3	10	13
5	2011-01-01	1	0	1	4	0	6	0	1	0.24	0.2879	0.75	0.0000	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17375	2012-12-31	1	1	12	19	0	1	1	2	0.26	0.2576	0.60	0.1642	11	108	119
17376	2012-12-31	1	1	12	20	0	1	1	2	0.26	0.2576	0.60	0.1642	8	81	89
17377	2012-12-31	1	1	12	21	0	1	1	1	0.26	0.2576	0.60	0.1642	7	83	90
17378	2012-12-31	1	1	12	22	0	1	1	1	0.26	0.2727	0.56	0.1343	13	48	61
17379	2012-12-31	1	1	12	23	0	1	1	1	0.26	0.2727	0.65	0.1343	12	37	49

17379 rows × 16 columns

Data Description

Numerical

- yr - Year (0: 2011, 1:2012)
- temp - Normalized temperature in Celsius.
- atemp - Normalized feeling temperature in Celsius.
- hum - Normalized humidity.
- windspeed - Normalized wind speed.
- casual - count of casual users
- registered – count of registered users
- cnt - count of total rental bikes including both casual and registered

Categorical

• season - Season (1:Winter, 2:Spring, 3:Summer, 4:Fall)
• mnth – Month (1 to 12)
• hr – Hour (0 to 23)
• holiday - whether day is holiday or not (extracted from Holiday Schedule)
• weekday – Day of the week
• workingday - If day is neither weekend nor holiday is 1, otherwise is 0
• weathersit - (extracted from Freemeteo)
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
• event – Data Outliers

Pre-Processing

Checking Missing Value

We can conclude that this dataset does not contain any missing values

	Column	DataType	Null	Null (%)	nUnique	Unique Samples
0	dteday	object	0	0.0	731	[2011-07-10, 2011-06-21]
1	season	int64	0	0.0	4	[4, 1]
2	yr	int64	0	0.0	2	[0, 1]
3	mnth	int64	0	0.0	12	[2, 1]
4	hr	int64	0	0.0	24	[10, 14]
5	holiday	int64	0	0.0	2	[1, 0]
6	weekday	int64	0	0.0	7	[0, 4]
7	workingday	int64	0	0.0	2	[1, 0]
8	weathersit	int64	0	0.0	4	[3, 1]
9	temp	float64	0	0.0	50	[0.34, 0.94]
10	atemp	float64	0	0.0	65	[0.8333, 0.9091]
11	hum	float64	0	0.0	89	[0.6, 0.41]
12	windspeed	float64	0	0.0	30	[0.2239, 0.5821]
13	casual	int64	0	0.0	322	[37, 72]
14	registered	int64	0	0.0	776	[383, 217]
15	cnt	int64	0	0.0	869	[354, 103]

Pre-Processing

Rescaling Procedure

- temp – The values are derived via (t-tmin)/(tmax-tmin), tmin=-8, t_max=+39 (only in hourly scale)
- atemp –The values are derived via (t-tmin)/(tmax-tmin), tmin=-16, t_max=+50 (only in hourly scale)
- hum – The values are divided to 100 (max)
- windspeed – The values are divided to 67 (max)

Renaming Value

- season

1 = Winter
2 = Spring
3 = Summer
4 = Fall

- yr

0 = 2011
1 = 2012

- weathersit

1 = Clear
2 = Cloudy
3 = Light Rain/Snow
4 = Heavy Rain/Snow

- workingday

1 = Work
0 = Non-Work

- weekday

0 = Sunday
1 = Monday
2 = Tuesday
3 = Wednesday
4 = Thursday
5 = Friday
6 = Saturday

- holiday

1 = Holiday
0 = Non-Holiday

- mnth

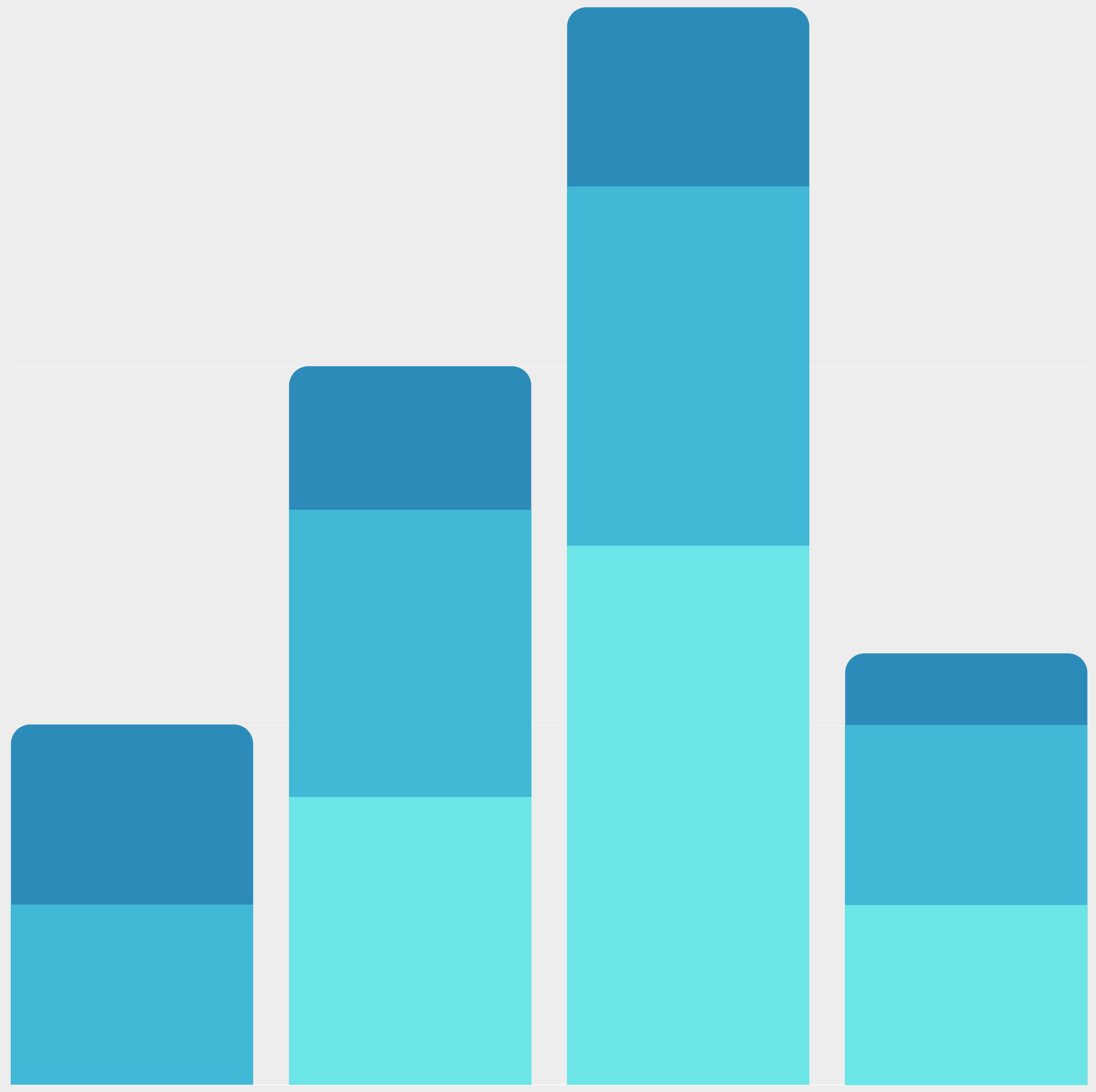
1 = January
2 = February
3 = March
4 = April
5 = May
6 = June
7 = July
8 = August
9 = September
10 = October
11 = November
12 = December

Descriptive Statistics

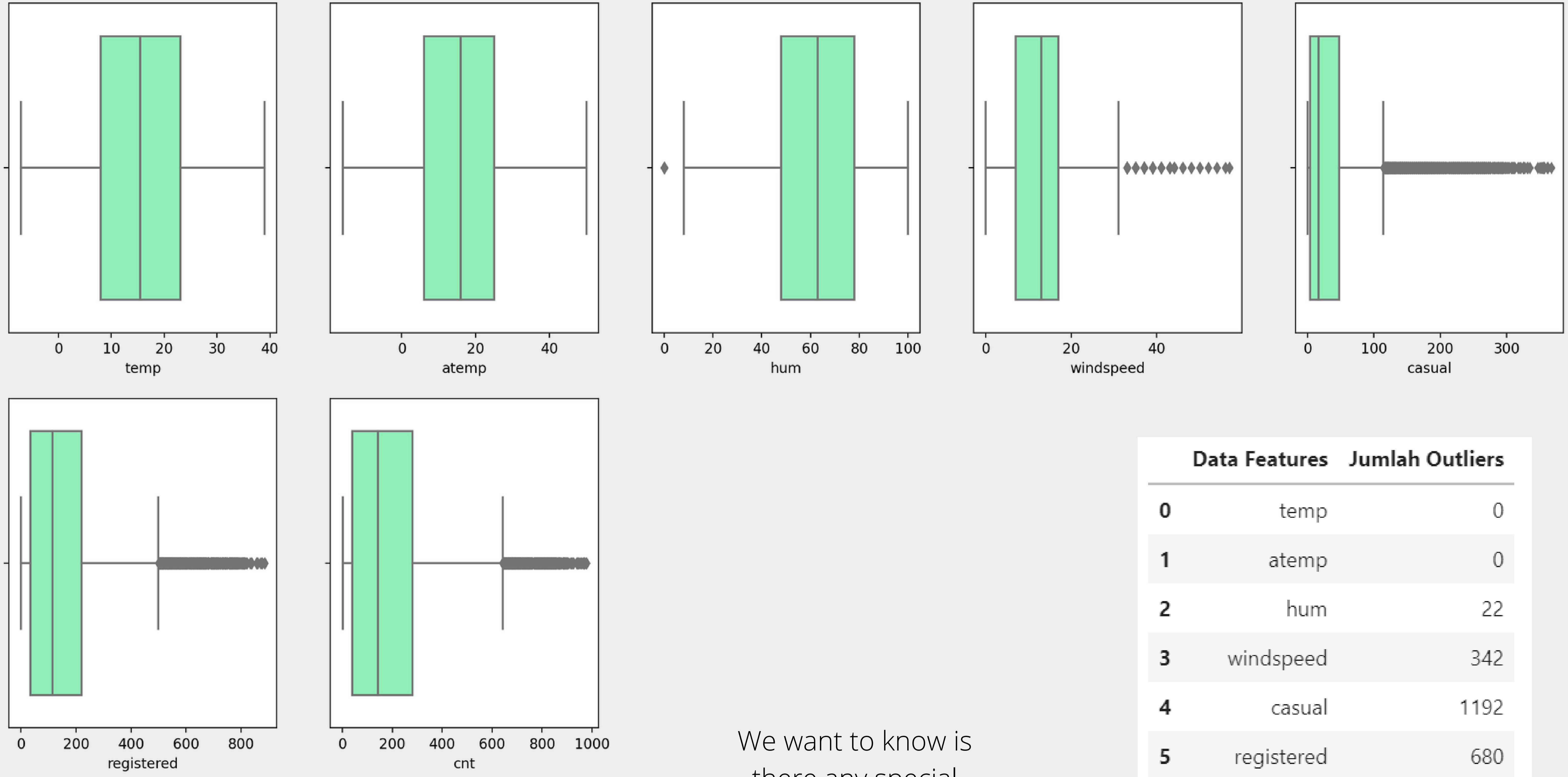
	temp	atemp	hum	windspeed	casual	registered	cnt
count	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000	17379.000000
mean	15.358397	15.401116	62.722884	12.736233	35.676218	153.786869	189.463088
std	9.050138	11.341858	19.292983	8.196891	49.305030	151.357286	181.387599
min	-7.060000	-16.000000	0.000000	0.000000	0.000000	0.000000	1.000000
25%	7.980000	6.000000	48.000000	7.000000	4.000000	34.000000	40.000000
50%	15.500000	16.000000	63.000000	13.000000	17.000000	115.000000	142.000000
75%	23.020000	25.000000	78.000000	17.000000	48.000000	220.000000	281.000000
max	39.000000	50.000000	100.000000	57.000000	367.000000	886.000000	977.000000

	season	yr	mnth	hr	holiday	weekday	workingday	weathersit
count	17379	17379	17379	17379	17379	17379	17379	17379
unique	4	2	12	24	2	7	2	4
top	Summer	2012	July	17	Non-Holiday	Saturday	Work	Clear
freq	4496	8734	1488	730	16879	2512	11865	11413

Data Exploration



Identifying Outliers



We want to know is there any special circumstances at this hours

	Data Features	Jumlah Outliers
0	temp	0
1	atemp	0
2	hum	22
3	windspeed	342
4	casual	1192
5	registered	680
6	cnt	505

Identifying Outliers (cont.)

Outliers happens at :

- Holiday : 2012-04-16 (DC Emancipation Day)
- Non-Holiday : (199 dates)
    a. 2011-08-23 - Virginia Earthquake
    b. 2012-04-29 - The USA Science & Engineering Festival
    c. 2012-09-09 - The 7th Annual Nations Triathlon
    d. etc

Insight and further action:

- Adjust bike storing based on the recurring events
- Adding new column 'event' for ML model

	Data Features	Jumlah Outliers
0	temp	0
1	atemp	0
2	hum	22
3	windspeed	342
4	casual	1192
5	registered	680
6	cnt	505

We want to know is there any special circumstances at this hours

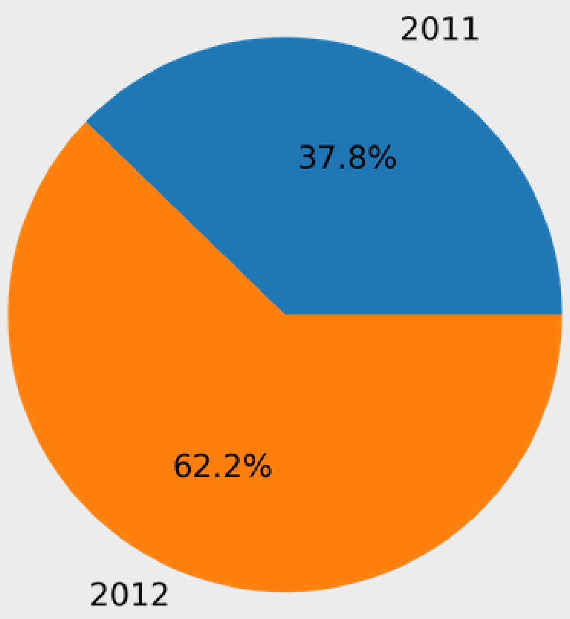
Identifying Outliers (cont.)

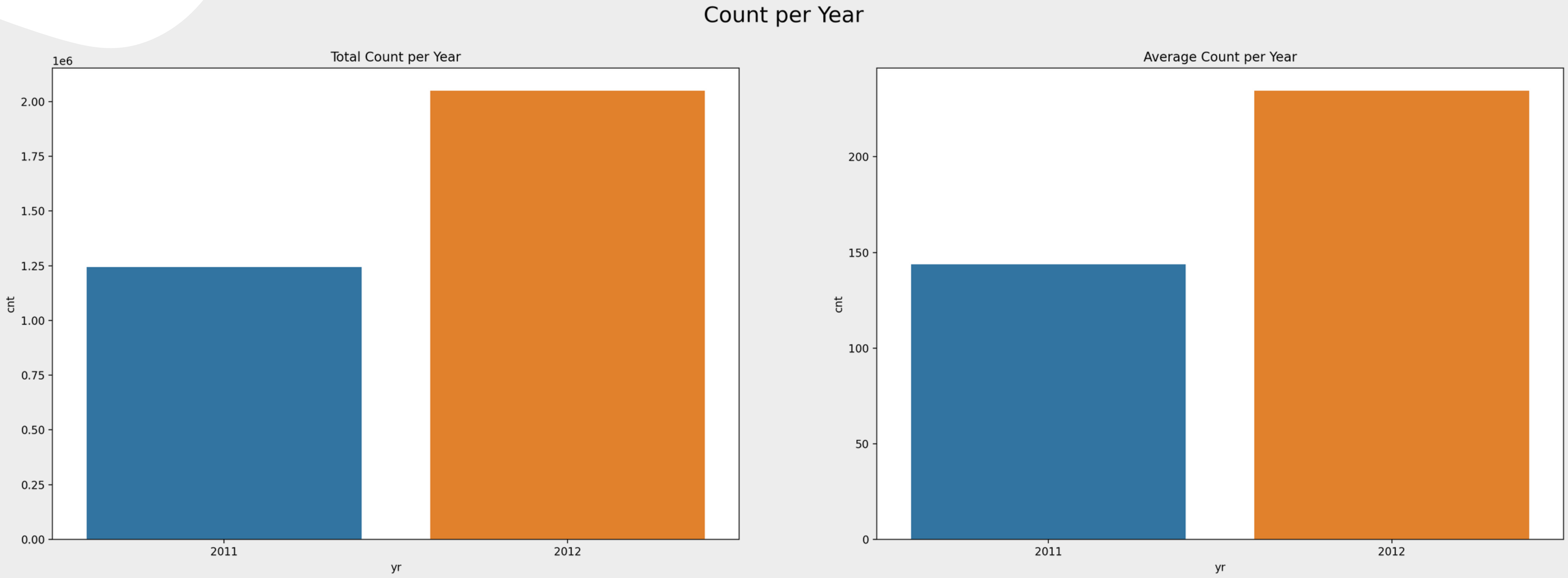
event column made using outliers data

	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	cnt	event
instant															
1	2011-01-01	Winter	2011	January	0	Non-Holiday	Saturday	Non-Work	Clear	3.28	3.0	81.0	0.0	16	No
2	2011-01-01	Winter	2011	January	1	Non-Holiday	Saturday	Non-Work	Clear	2.34	2.0	80.0	0.0	40	No
3	2011-01-01	Winter	2011	January	2	Non-Holiday	Saturday	Non-Work	Clear	2.34	2.0	80.0	0.0	32	No
4	2011-01-01	Winter	2011	January	3	Non-Holiday	Saturday	Non-Work	Clear	3.28	3.0	75.0	0.0	13	No
5	2011-01-01	Winter	2011	January	4	Non-Holiday	Saturday	Non-Work	Clear	3.28	3.0	75.0	0.0	1	No

Growth from 2011–2012

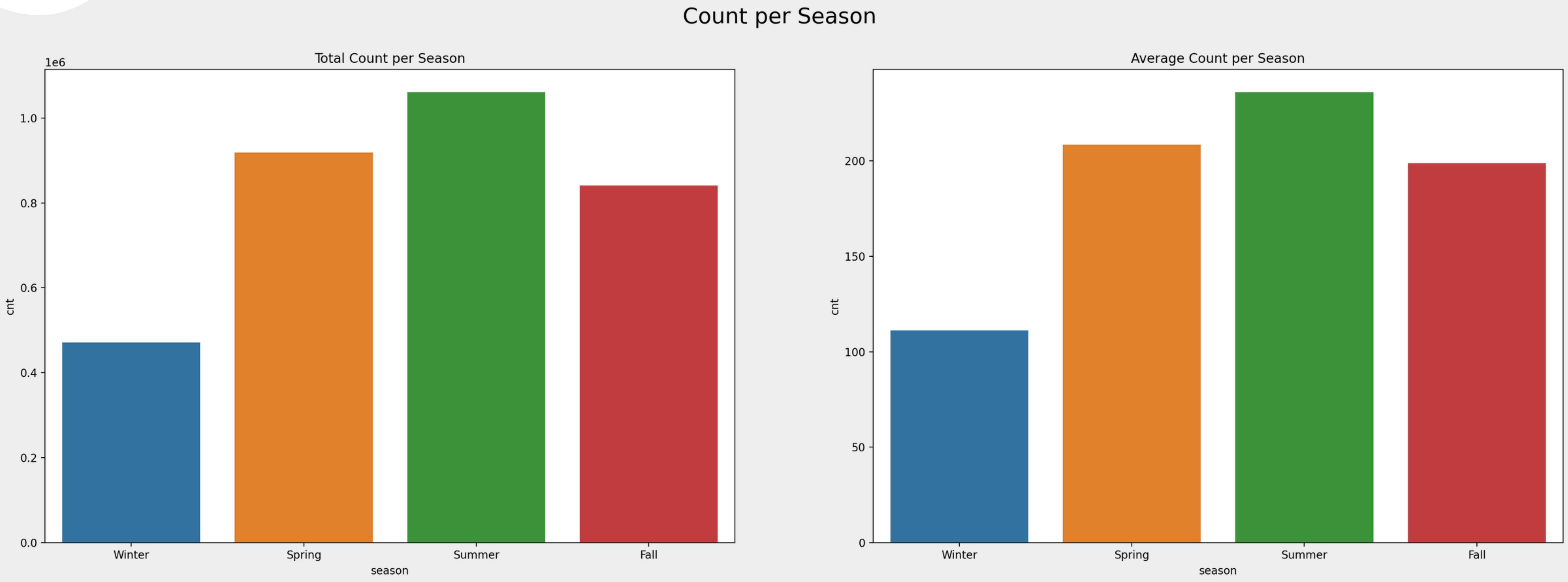
- Decent growth just over a year, 64.87%
- We expect increasing trends in the following years ahead





Distribution over Seasons

- Popular time (seasons) are Spring and Summer
- Can be used as a rough guide for the bike storing management system and promotion

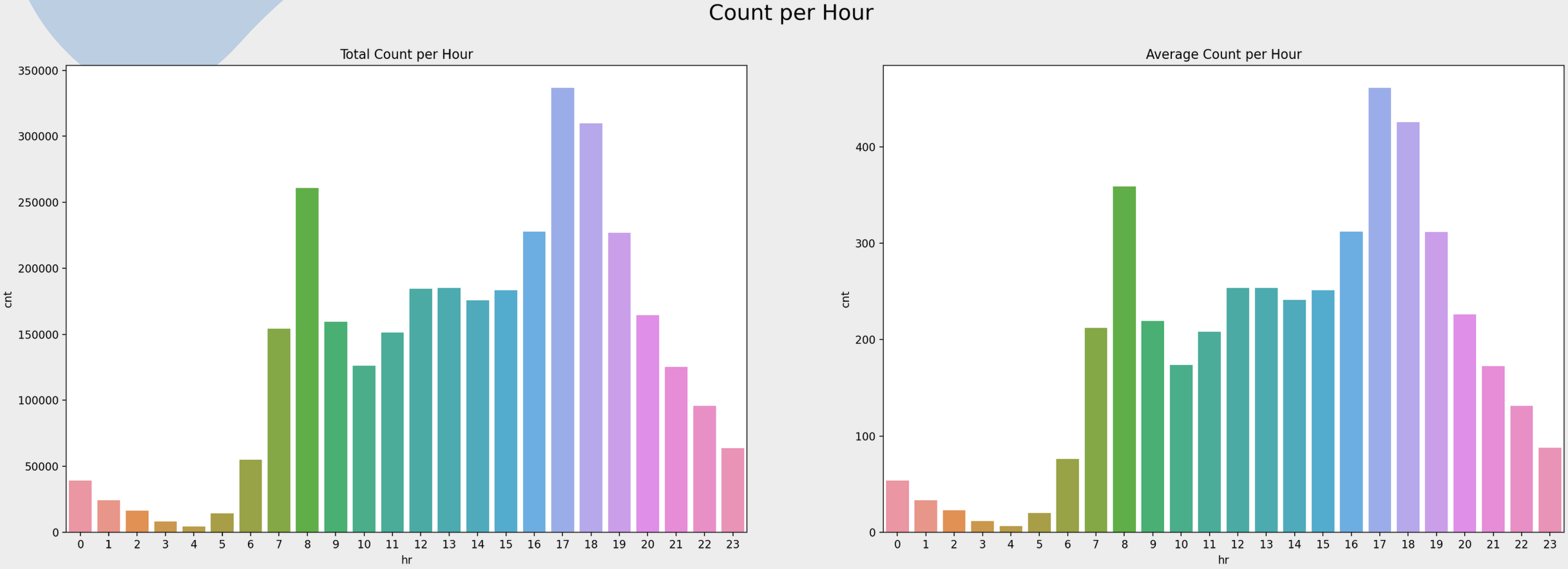


Distribution over hour

We want to determine the right time to do maintenance to minimize customer loss

The following is the general distribution regardless the day:

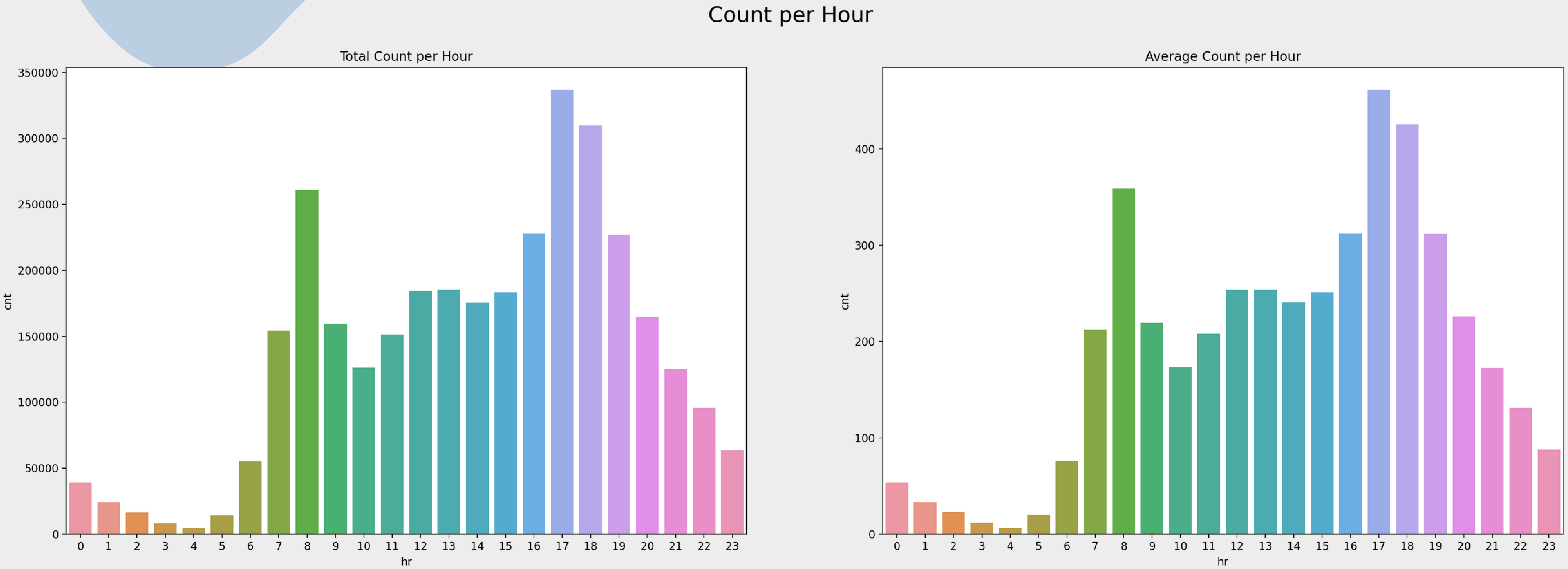
- Peaks : 8 am, 5 pm
- Valley : 4 am



Distribution over hour (Cont.)

Are these peaks/valleys stable?

.. we need to vary other features as well to answer this..



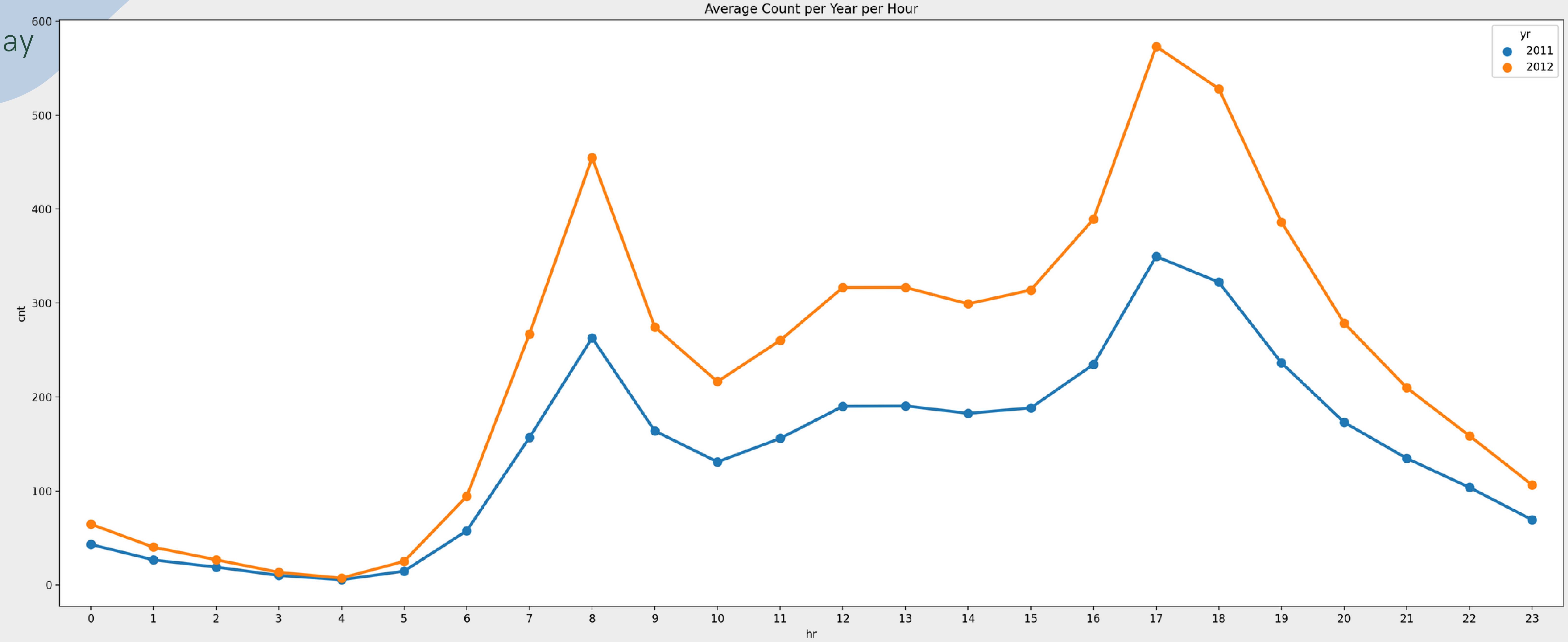
Varying other features to see if the pattern is stable

Varying features:

- yr
- season
- weekday
- workingday

Are these peaks/valleys stable over years?

Yes



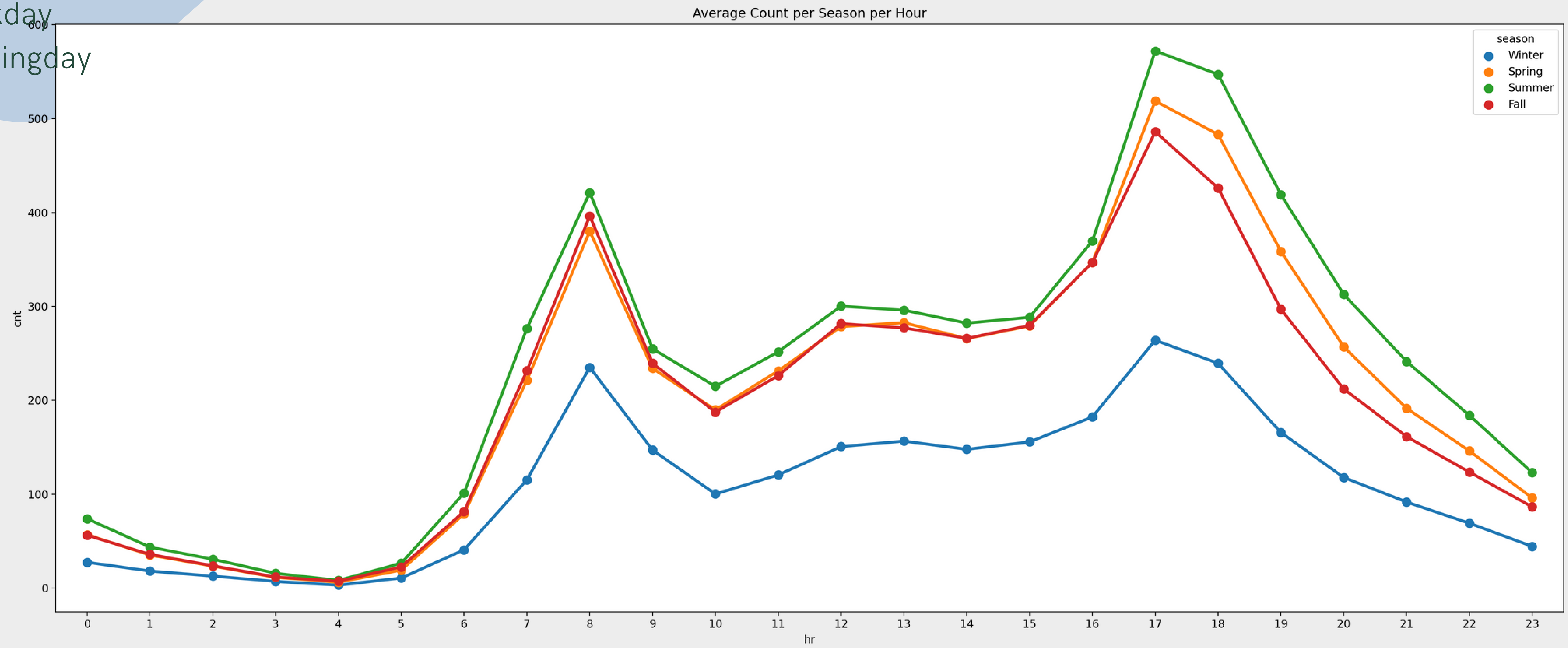
Varying other features to see if the pattern is stable

Varying features:

- yr
- season
- weekday
- workingday

Are these peaks/valleys stable over season?

Yes



Varying other features to see if the pattern is stable

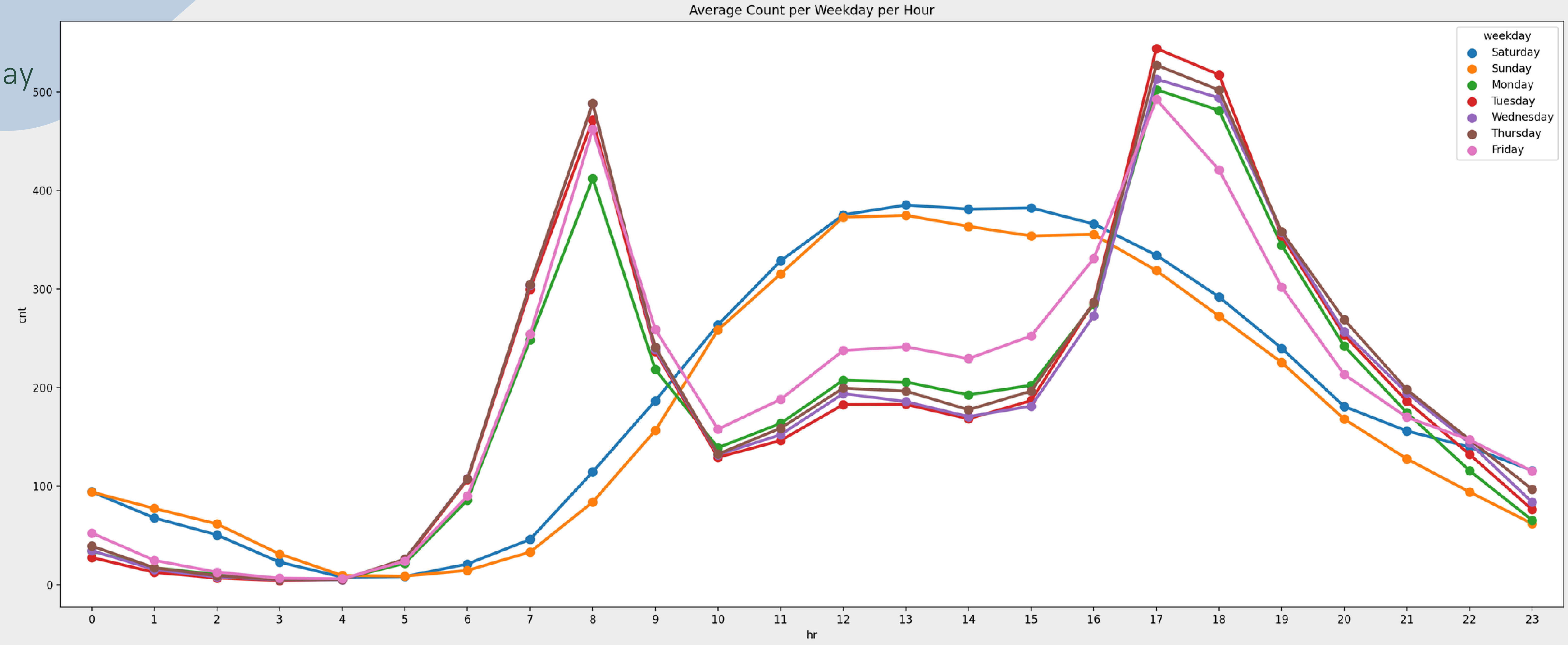
Varying features:

- yr
- season
- weekday
- workingday

Are these peaks/valleys stable over weekday?

Yes. But different pattern appear on weekend

Guess?



Varying other features to see if the pattern is stable

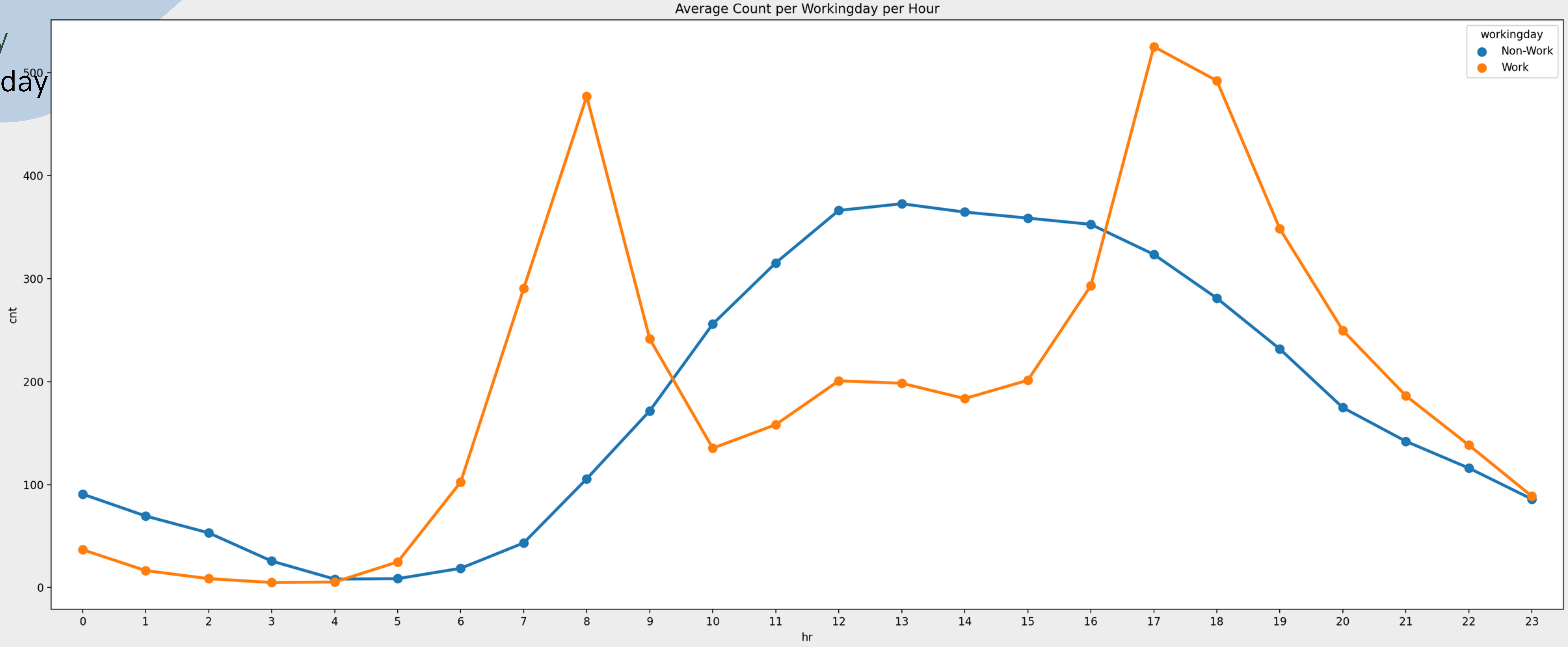
Varying features:

- yr
- season
- weekday
- workingday

Are these peaks/valleys stable over workingday?

Yes. But different pattern appear on weekend

Guess? Different patterns for working/non-working days



Insights and Recommendations

Bike maintenance/storing recommendations :

- Daily maintenance (light maintenance) can be done around 4 am.
- Weekly maintenance (routine maintenance) recommended to be done on Sunday.
- Regular events, such as holidays, should be taken into consideration into bike maintenance/storing schedule.
- On bad weather, e.g. winter, unused bikes should be stored in the warehouse to reduce destructive effect of weather to the bikes.

Insights and Recommendations (cont.)

More on recommendations :

Based on research, 90% of bike usage is within 30-60 minutes, so we can have the following storing/maintenance scenario as follows:

- Total bikes in station at 5 am is 20
- Total bikes needed at 8 am is 360
- Total bikes needed at 5 pm is 450
- Total bikes needed at 8 pm is 220

We want to do storing-pickup twice a day,

- one for adding bikes to station
- one for collect bikes from station

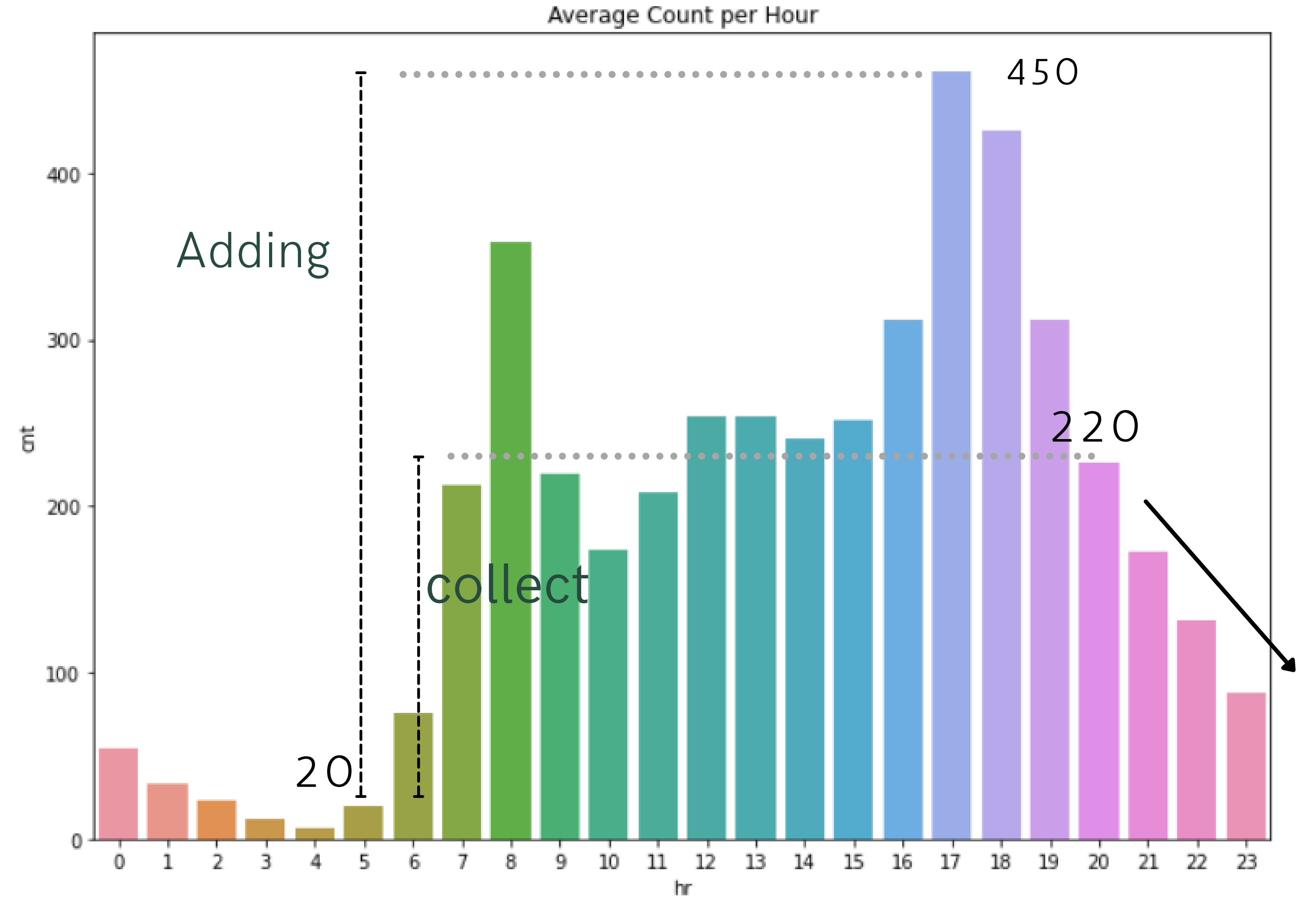
So, at 5 am we add

#bikes needed at 5 pm - #bikes in station at 5 am = 450-20 = 430

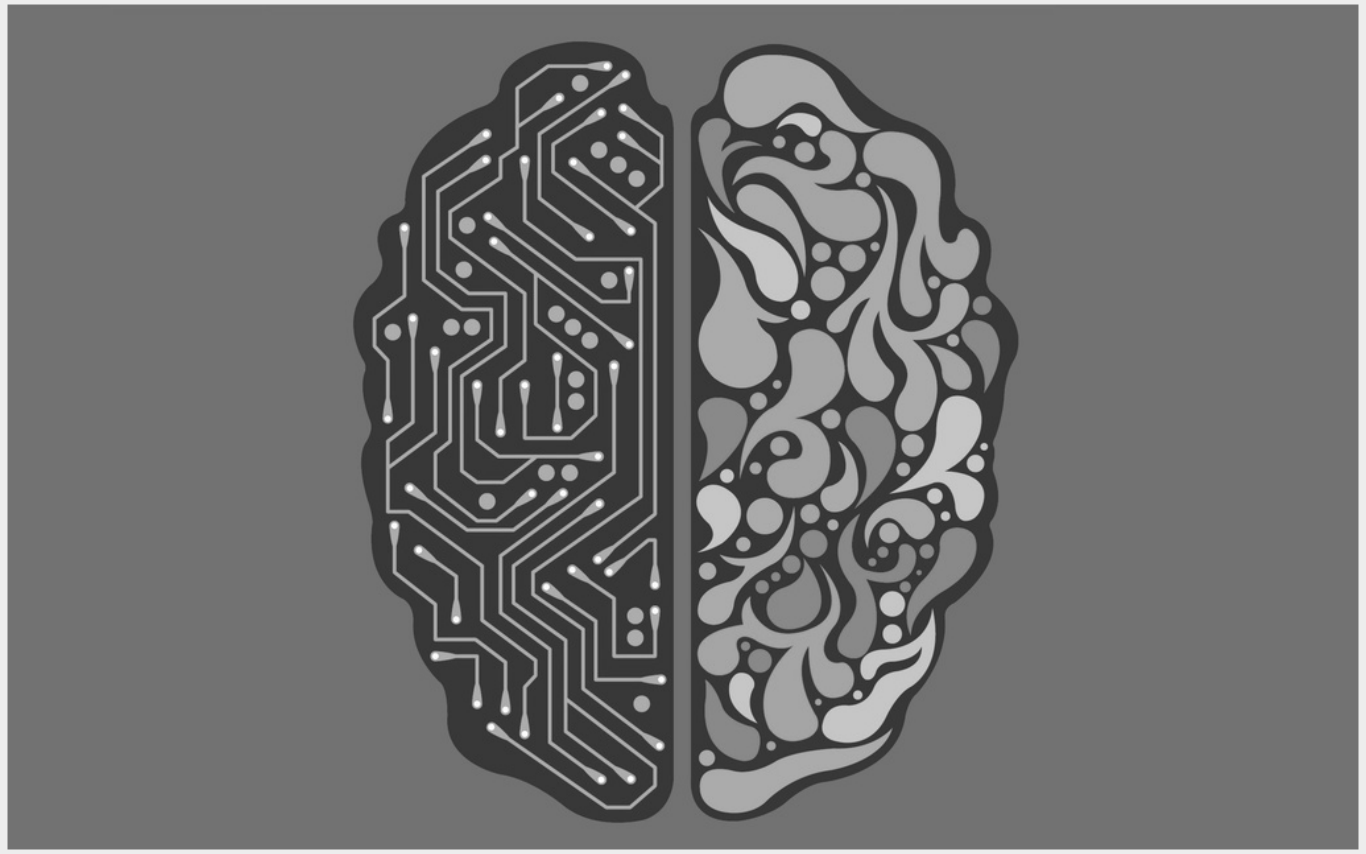
and at 8 pm we collect

#bikes in station at 5 pm - #bikes needed at 8 pm = 450-220=230

Insights and Recommendations (cont.)



Bike Demand Prediction using Machine Learning



Data Preparation

- Feature Selection
- Splitting Data

Training Model

- Decision Tree Regressor
- Random Forest Regressor
- XGB Regressor

Evaluating Model

- R-Squared
- Median Absolute Error
- Residual Analysis

Improving Model

Hyperparameter Tuning with

- Grid Search
- Randomized Search
- Bayesian Search

Machine Learning - Regression

Predicting bike share demand (*cnt* column) using regression

Data Preparation

Feature Selection

- 'dteday' (dropped)
- 'season'
- 'yr'
- 'mnth'
- 'hr'
- 'holiday'
- 'weekday'
- 'workingday'
- 'weathersit'
- 'atemp'
- 'temp' (dropped)
- 'hum'
- 'windspeed'
- 'casual' (dropped)
- 'registered'(dropped)
- 'event'

With 'cnt' as Target Column

Splitting Data

Splitting proportion for train and test as

- Train 75%
- Test 25%

Training & Evaluating Base Models Data Testing

Evaluation Matrix Comparison

	index	R Squared	MAE	MSE	RMSE	MedAE	Residual_Negative
0	Base DecisionTree	90.023868	34.598964	3308.480380	57.519391	19.000000	0.878532
1	Base RandomForest	94.908314	25.702521	1688.604542	41.092634	14.980000	0.530468
2	Base XGBoost	94.987489	26.070419	1662.347040	40.771890	15.494263	0.504172

From the table above, we already get good evaluation matrix score on base models, but we will try to do some parameter tuning to see possibilities of improvement.

Hyperparameter Tuning

Decision Tree

Best parameters

- Bayesian Search
  - max_depth = 49
  - max_features = auto
  - min_samples_leaf = 1
  - min_samples_split = 16
- Grid Search 1
  - max_depth = 15
  - max_features = None
  - min_samples_leaf = 2
  - min_samples_split = 12
- Grid Search 2
  - criterion = 'mae'
  - max_depth = 17
  - max_features = None
  - min_samples_leaf = 2
  - min_samples_split = 16

Random Forests

Best parameters

- Bayesian Search
  - max_depth = 76
  - max_features = auto
  - min_samples_leaf = 2
  - min_samples_split = 2
  - n_estimators = 6800
- Randomized Search
  - n_estimators = 6170
  - min_samples_split = 2
  - min_samples_leaf = 1
  - max_features = auto
  - max_depth = 26

XGBoosts

Best parameters

- Bayesian Search
  - learning_rate = 0.1
  - max_depth = 6
  - n_estimators = 1600
- Randomized Search
  - n_estimators=2300
  - max_depth=31
  - learning_rate=0.3

Choosing The Best Model

Evaluation Matrix Comparison

	index	R Squared
0	Base DecisionTree	90.023868
1	Base RandomForest	94.908314
2	Base XGBoost	94.987489
3	RandomForest Tuned Bayes	95.029671
4	RandomForest Tuned Random	94.997472
5	DecisionTree Tuned Bayes	92.233244
6	DecisionTree Tuned Grid 1	92.081840
7	DecisionTree Tuned Grid 2	91.192706
8	XGBoost Tuned Random	94.824871
9	XGBoost Tuned Bayes	95.553821

Evaluation Matrix Comparison

	index	R Squared	MAE	MSE	RMSE	MedAE	Residual_Negative
0	Base DecisionTree	90.023868	34.598964	3308.480380	57.519391	19.000000	0.878532
1	Base RandomForest	94.908314	25.702521	1688.604542	41.092634	14.980000	0.530468
2	Base XGBoost	94.987489	26.070419	1662.347040	40.771890	15.494263	0.504172
3	RandomForest Tuned Bayes	95.029671	25.373437	[illegible]	[illegible]	[illegible]	[illegible]
4	RandomForest Tuned Random	94.997472	25.494080	1659.036404	40.731271	[illegible]	0.529547
5	DecisionTree Tuned Bayes	92.233244	31.268196	2575.763885	50.751984	17.250000	0.514471
6	DecisionTree Tuned Grid 1	92.081840	31.551836	2625.975484	51.244273	17.500000	0.511997
7	DecisionTree Tuned Grid 2	91.192706	32.342923	2920.847411	54.044865	18.000000	0.543472
8	XGBoost Tuned Random	94.824871	25.883868	1716.277645	41.427981	14.959915	0.538351
9	XGBoost Tuned Bayes	95.553821	24.242375	1474.529058	38.399597	14.732788	0.501582

Best model that we choose is RandomForest Tuned Random, because it’s R-Squared is 94.99, it’s MedAE is lower than the average of all MedAE and Residual Negative is 52.95%.

Residual Plot

Assumptions

From Capital-Bikeshare data, bike rent fee is $2 per 30 minute (or $4 per hour).

- Assume from $4, $2 allocated for maintenance cost per bike.
- Assume previously the company used conventional method of predicting bike demand :
  - By taking average or
  - Q3 from the previous month to predict the bike-demand.

So if

y_pred < y_true then we might lost customer (lost $4 times y_true - y_pred)

whereas if

y_pred > y_true then we might lost $2 for maintenance for each unused bikes

Based on these assumptions, we want to compare the monthly cost by machine learning model and conventional method

Simple Model Simulation

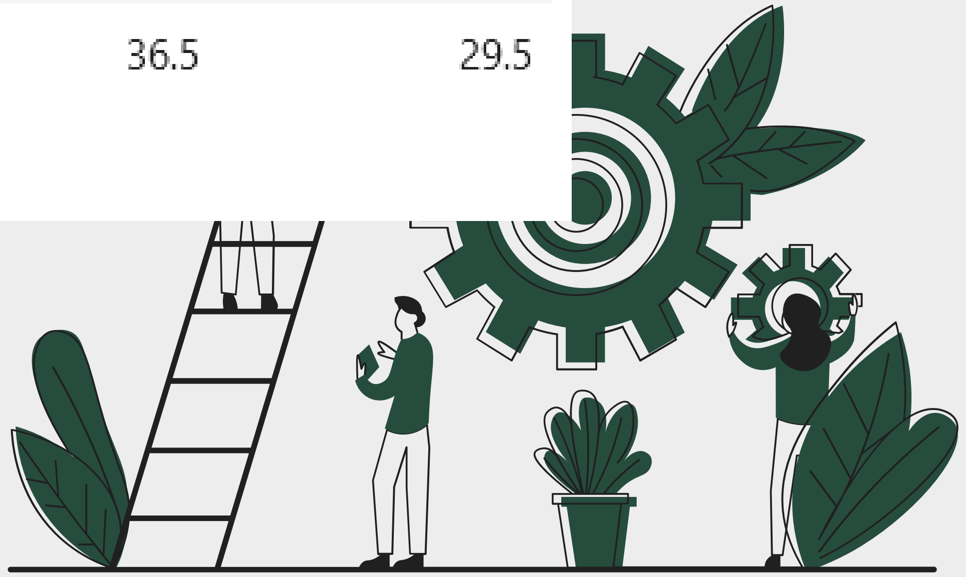
Simple Model Simulation (Cont.)

Comparing cost prediction for a month (December 2012)

	cnt	predRF_Random	predQ3	predMean	gap_predRF_Random	gap_predQ3	gap_predMean	sales	cost_predRF_Random	cost_predQ3	cost_predMean
instant											
16638	108	115.0	111.50	94.50	-7.0	-3.50	13.50	432	14.0	7.0	54.0
16639	69	97.0	87.50	68.00	-28.0	-18.50	1.00	276	56.0	37.0	4.0
16640	50	59.0	46.75	42.75	-9.0	3.25	7.25	200	18.0	13.0	29.0
16641	15	16.0	22.75	20.00	-1.0	-7.75	-5.00	60	2.0	15.5	10.0
16642	5	10.0	9.25	8.00	-5.0	-4.25	-3.00	20	10.0	8.5	6.0
...	...	...	...	...	...	...	...	...	...	...	...
17375	119	290.0	380.50	375.50	-171.0	-261.50	-256.50	476	342.0	523.0	513.0
17376	89	202.0	283.75	270.25	-113.0	-194.75	-181.25	356	226.0	389.5	362.5
17377	90	139.0	186.50	172.75	-49.0	-96.50	-82.75	360	98.0	193.0	165.5
17378	61	110.0	113.50	111.00	-49.0	-52.50	-50.00	244	98.0	105.0	100.0
17379	49	60.0	67.25	63.75	-11.0	-18.25	-14.75	196	22.0	36.5	29.5

742 rows × 11 columns

Example simulation for RandomForest Tuned Randomized Search



Simple Model Simulation (Cont.)

Comparison of Actual Sales and Cost of each Method/Model

sales	494852.000000
cost_predDT	127782.000000
cost_predRF	93356.000000
cost_predXGB	90902.000000
cost_predDT_Bayes	119728.000000
cost_predDT_Grid1	119176.000000
cost_predDT_Grid2	111080.000000
cost_predRF_Bayes	90514.000000
cost_predRF_Random	92460.000000
cost_predXGB_Random	108874.000000
cost_predXGB_Bayes	87686.000000
cost_predQ3	130795.000000
cost_predMean	123343.733333

Percentage of Cost to Actual Sales

cost_predDT	25.822266
cost_predRF	18.865439
cost_predXGB	18.369533
cost_predDT_Bayes	24.194709
cost_predDT_Grid1	24.083160
cost_predDT_Grid2	22.447116
cost_predRF_Bayes	18.291125
cost_predRF_Random	18.684374
cost_predXGB_Random	22.001326
cost_predXGB_Bayes	17.719641
cost_predQ3	26.431135
cost_predMean	24.925378

- Even though Machine Learning Model is not perfect yet, but this Model have lower 6%-8% per month Cost than conventional method.

Future Works

To conclude, We understand that Our Machine Learning Prediction on this Project is not perfect. Because there's several factors that can improve Our Model Prediction such as:

- Cost/Sales of the Bike Rent
- Duration of Bike Usage
- Distance of Bike Usage
- Numbers of Bike Station and their Specific Location

We can try to define our custom metric performance that has more related to our business goals as follows

$$MAE_{\text{new}} = \alpha \frac{\sum_{i=1}^{n} |y_i - y_i^{pred}|_{\text{pos}}}{n} + \beta \frac{\sum_{i=1}^{n} |y_i - y_i^{pred}|_{\text{neg}}}{n}, \quad \alpha > 1 \text{ dan } 0 < \beta \leq 1$$

y_i : nilai aktual

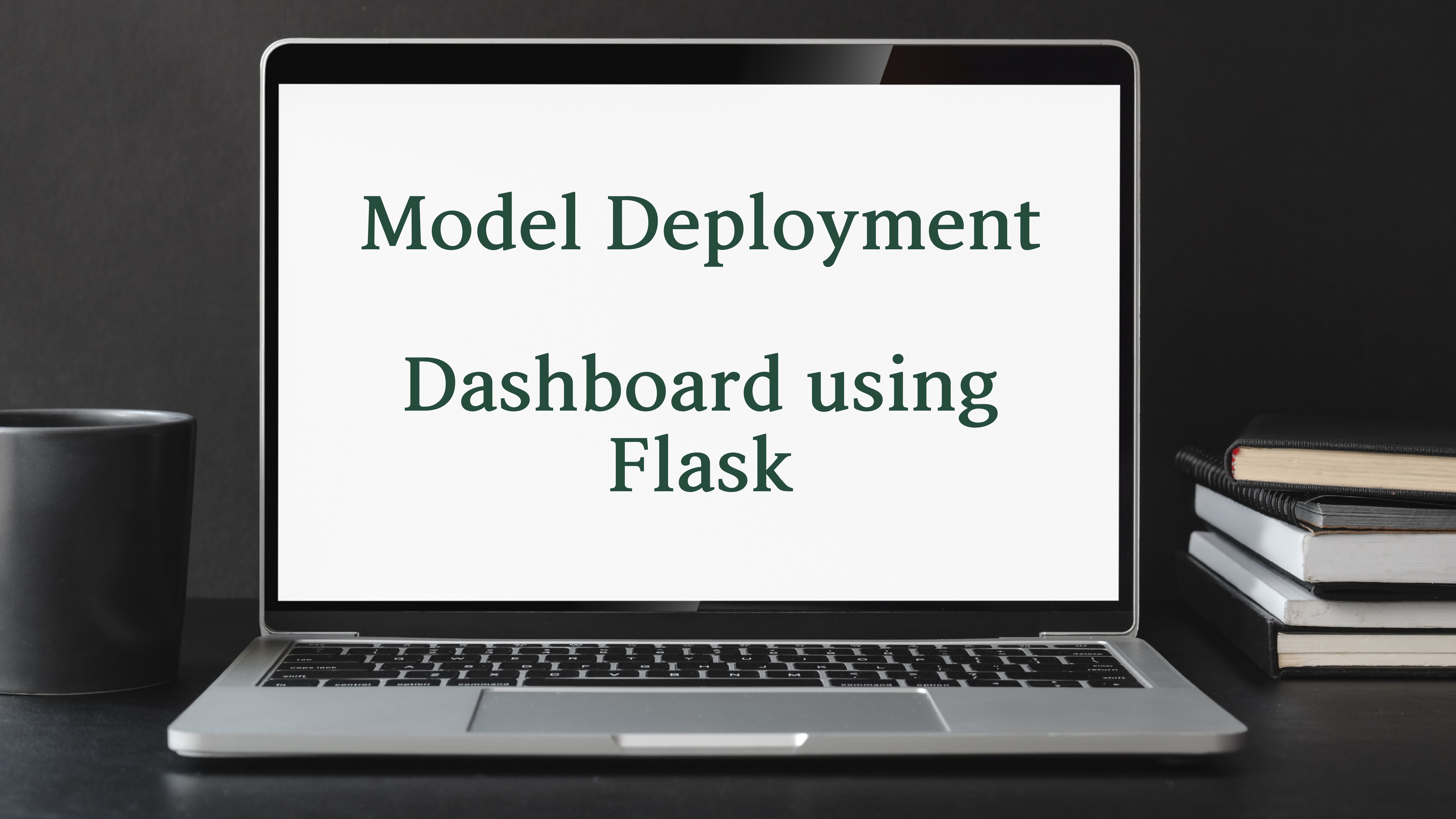
y_i^{pred} : nilai prediksi dari model

n : jumlah data testing

α : bobot penalti untuk residu positif (residu positif tidak diinginkan sehingga diberikan penalti besar)

β : bobot penalti untuk residu negatif (residu negatif lebih diinginkan sehingga diberikan penalti yang lebih kecil)

we can try to minimize this metric when we use gridsearch or other searching method.


Model Deployment
Dashboard using
Flask

Model Deployment Preview

Bike Sharing in Washington D.C. Dataset

	season	year	month	hour	holiday	day	work	weather	atmep	hum	wind	cnt	event
*	Winter	2011	January	0	Non-Holiday	Saturday	Non-Work	Clear	3.0	81.0	0.0	16	No
*	Winter	2011	January	1	Non-Holiday	Saturday	Non-Work	Clear	2.0	80.0	0.0	40	No
*	Winter	2011	January	2	Non-Holiday	Saturday	Non-Work	Clear	2.0	80.0	0.0	32	No
*	Winter	2011	January	3	Non-Holiday	Saturday	Non-Work	Clear	3.0	75.0	0.0	13	No
*	Winter	2011	January	4	Non-Holiday	Saturday	Non-Work	Clear	3.0	75.0	0.0	1	No
*	Winter	2011	January	5	Non-Holiday	Saturday	Non-Work	Cloudy	1.0	75.0	6.0	1	No
*	Winter	2011	January	6	Non-Holiday	Saturday	Non-Work	Clear	2.0	80.0	0.0	2	No
*	Winter	2011	January	7	Non-Holiday	Saturday	Non-Work	Clear	1.0	86.0	0.0	3	No
*	Winter	2011	January	8	Non-Holiday	Saturday	Non-Work	Clear	3.0	75.0	0.0	8	No

Bike Prediction Washington DC

Season :	Season ?
Year :	
Month :	Month ?
Hour :	Hour ?
Holiday :	Holiday ?
Day :	Day ?
Workingday :	Workingday ?
Weather :	Weather ?
Feels-Like-Temperature :	
Humidity :	
Windspeed :	
Event :	Event ?

Predict Bike

Back to Home

Bike Prediction Washington DC

Season: Spring
Year: 2021
Month: June
Hour: 14
Holiday: Non-Holiday
Day: Monday
Workingday: Work
Weather: Clear
Feels-Like-Temperature: 35.0
Humidity: 59.0
Windspeed: 20.0
Event: No
Bike Prediction : 261.0

Predict Again   Back to Home

Reference Links

- https://www.capitalbikeshare.com/
- https://dchr.dc.gov/page/holiday-schedules-2010-and-2011
- https://dchr.dc.gov/page/holiday-schedules-2012-and-2013
- http://www.freemeteo.com
- https://www.weather.gov/jetstream/beaufort_max
- https://www.bicyclehabitat.com/how-to/a-simple-bike-maintenance-chart-pg366.htm
- http://mobility-workspace.eu/wp-content/uploads/03_Working_Paper_Costs_October_2013.pdf
- https://www.bikenbike.co.id/sepeda-bnb-11/#:~:text=Alloy%20memiliki%20umur%20yang%20lebih,itu%20kekuatan%20akan%20terus%20menurun
- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.make_scorer.html

Thank you!

PURWADHIKA JCDS12

SCIKIT-LEARN - TEAM

https://github.com/PurwadhikaDev/ScikitLearnGroup_JC_DS_12_FinalProject