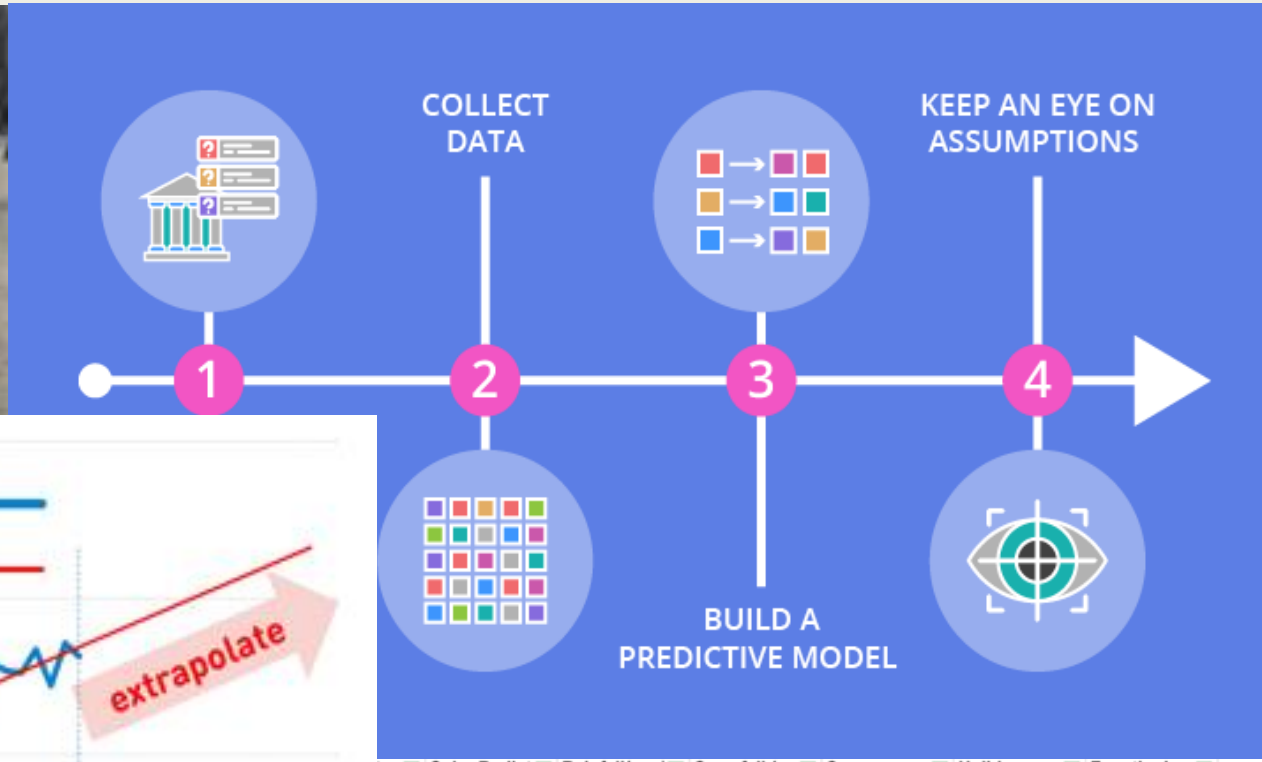
Seoul Bike Sharing Demand Forecasting

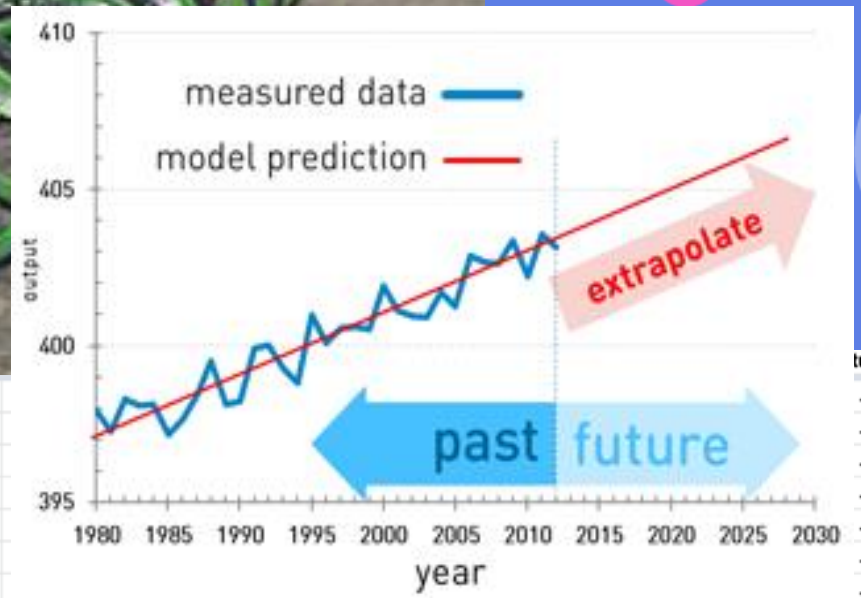
Ajith Kumar K
Biswarup Das Sarma
Ganesh Halthota
Girish Naik
Harsha Pamarthi Vardhan
Jamsheed MP
Shilpa Singh
Sourabh Gothe Vasant

Indian Institute of Science

DA-224-O MP1

Description of scope





0	01/12/17	254	0					-17.6	0	0	0	Winter	No Holiday	Yes		
1	01/12/17	204	1					-17.6	0	0	0	Winter	No Holiday	Yes		
2	01/12/17	173	2					-17.7	0	0	0	Winter	No Holiday	Yes		
3	01/12/17	107	3					-17.6	0	0	0	Winter	No Holiday	Yes		
4	01/12/17	78	4					-18.6	0	0	0	Winter	No Holiday	Yes		
5	01/12/17	100	5					-18.7	0	0	0	Winter	No Holiday	Yes		
6	01/12/17	181	6					-19.5	0	0	0	Winter	No Holiday	Yes		
7	01/12/17	460	7	-7.4	38	0.9	2000	-19.3	0	0	0	Winter	No Holiday	Yes		
8	01/12/17	930	8	-7.6	37	1.1	2000	-19.8	0.01	0	0	Winter	No Holiday	Yes		
9	01/12/17	490	9		27	0.5	1928	-22.4	0.23	0	0	Winter	No Holiday	Yes		
10	01/12/17	339	10	-3.5	24	1.2	1996	-21.2	0.65	0	0	Winter	No Holiday	Yes		

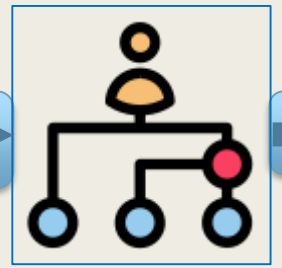
img src: https://english.seoul.go.kr/wp-content/uploads/2017/03/bike_2.jpg
https://www.dexlabanalytics.com/wp-content/uploads/2018/04/predictive_analytics_and_cross-selling-01_1.png
https://www.dexlabanalytics.com/wp-content/uploads/2018/04/large.png

E2E Process Flow

Datatype conversion
Histograms

Trend Analysis
Correlation Analysis

Exploratory Data Analysis



Time
Seasons

Stratification

Train-Test Split

Pipeline

Simple Imputer
Linear Regression
KN Regressor
Bfill, Ffill

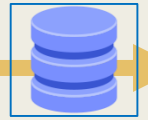
Imputation

OHE
Standard Scalar
Cyclic Trigonometry
Polynomial regression

Feature Engineering

Test Data 20%

Train Data 80%



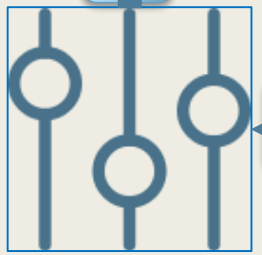
Iterations

R2
RMSE

Prediction Evaluation

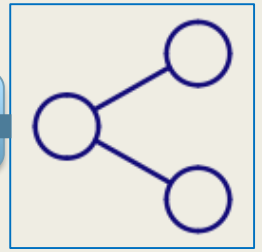
Linear Regression
Ridge
Lasso
Elastic net
KN Regression
XGBoost

ML Model

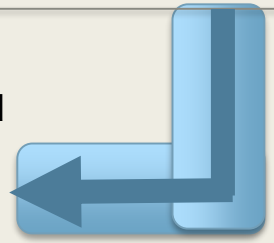


K-Fold

Cross-Validations

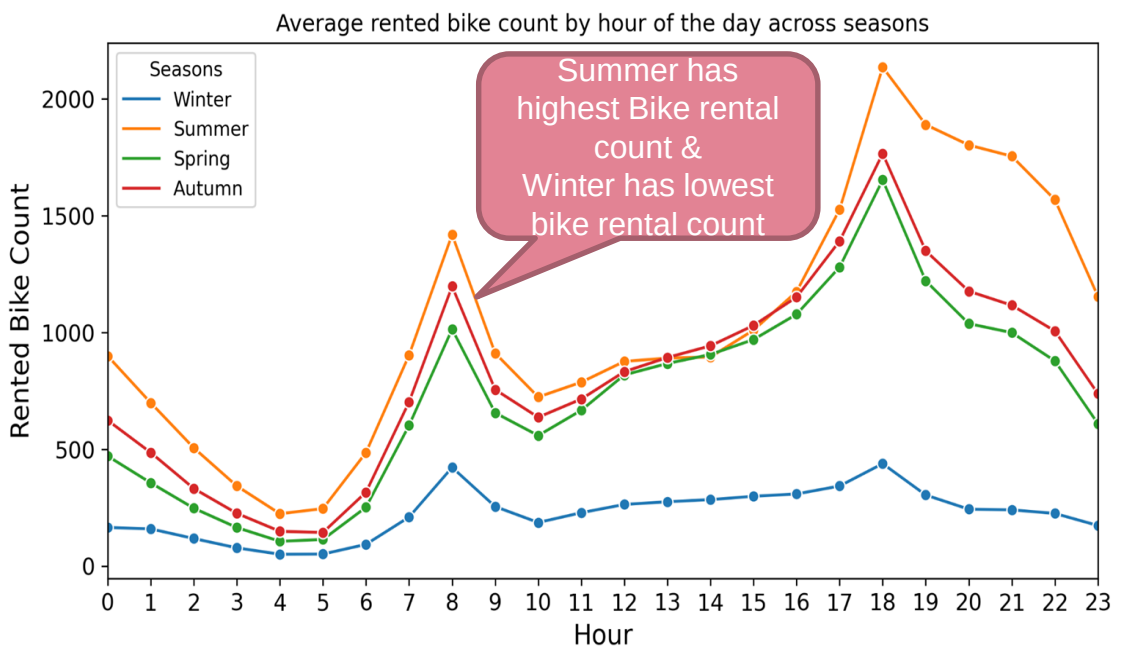


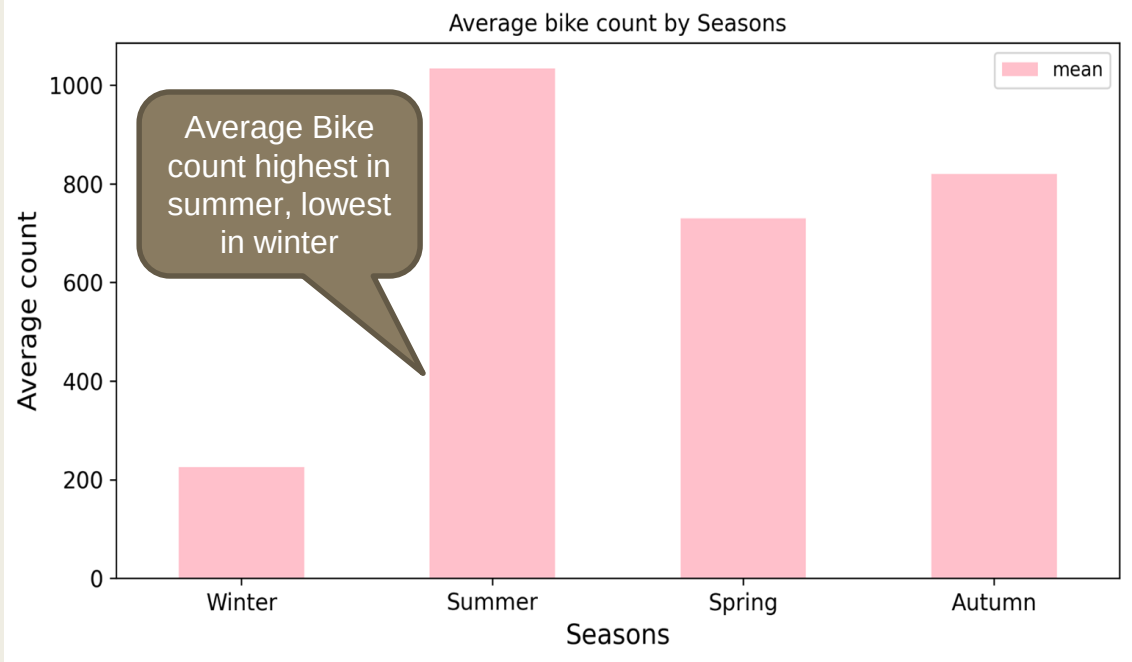
Seasons-based
Precipitation
Day-Night
Time-based
Holiday/No Holiday

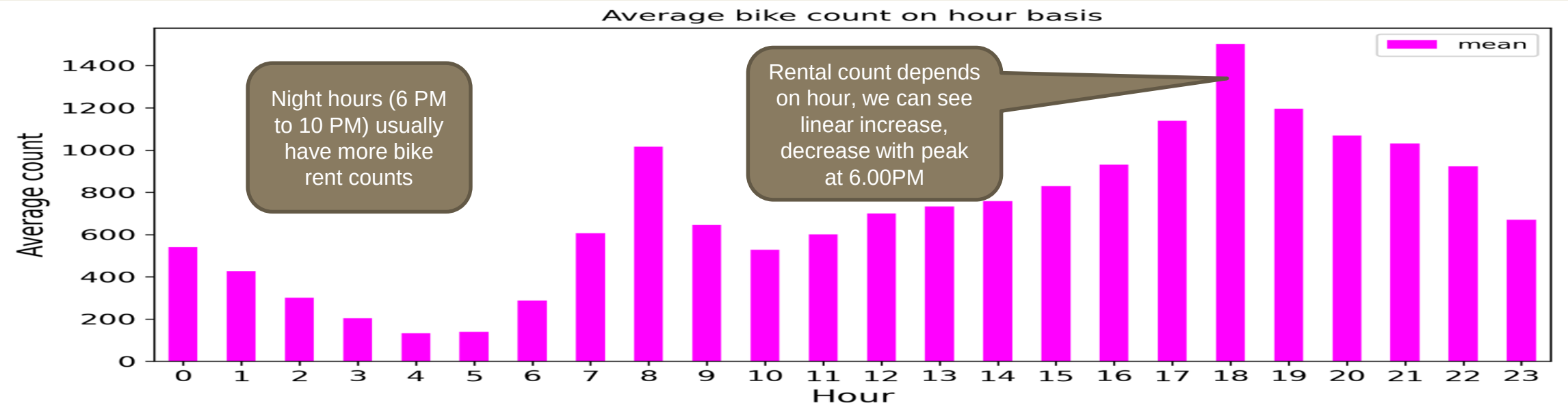


Data Split for Ensemble Modelling

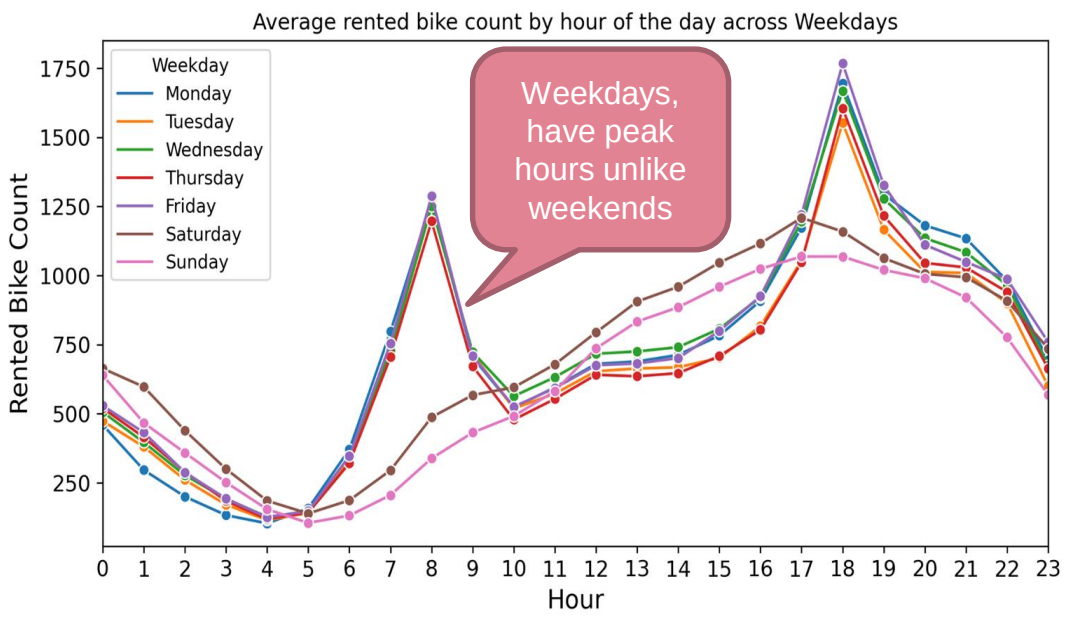
Exploratory data analysis

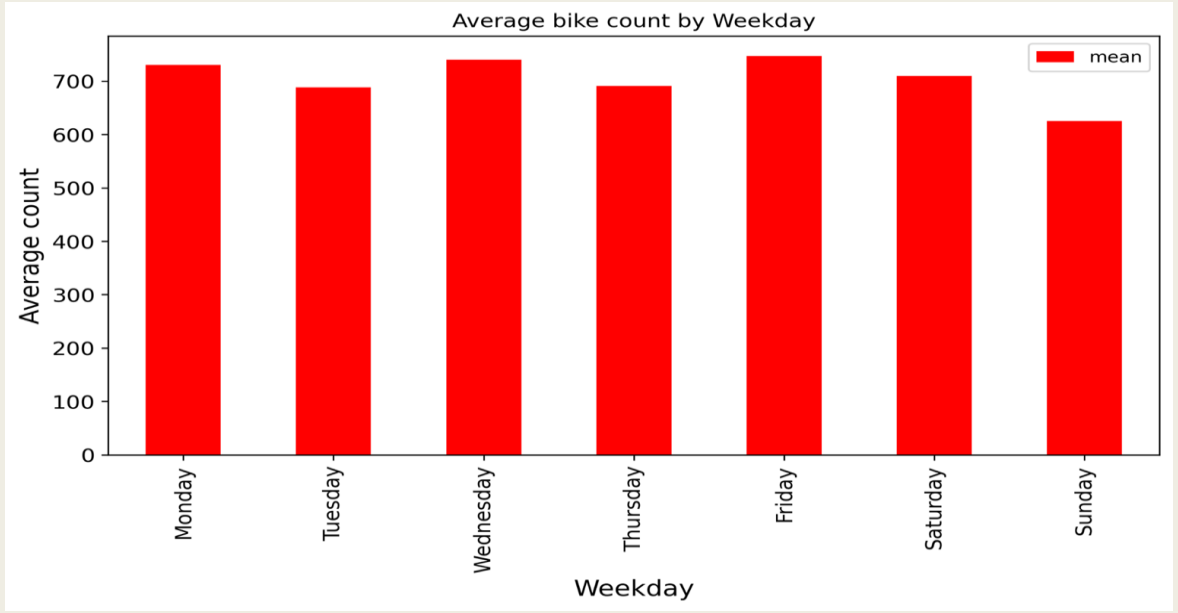


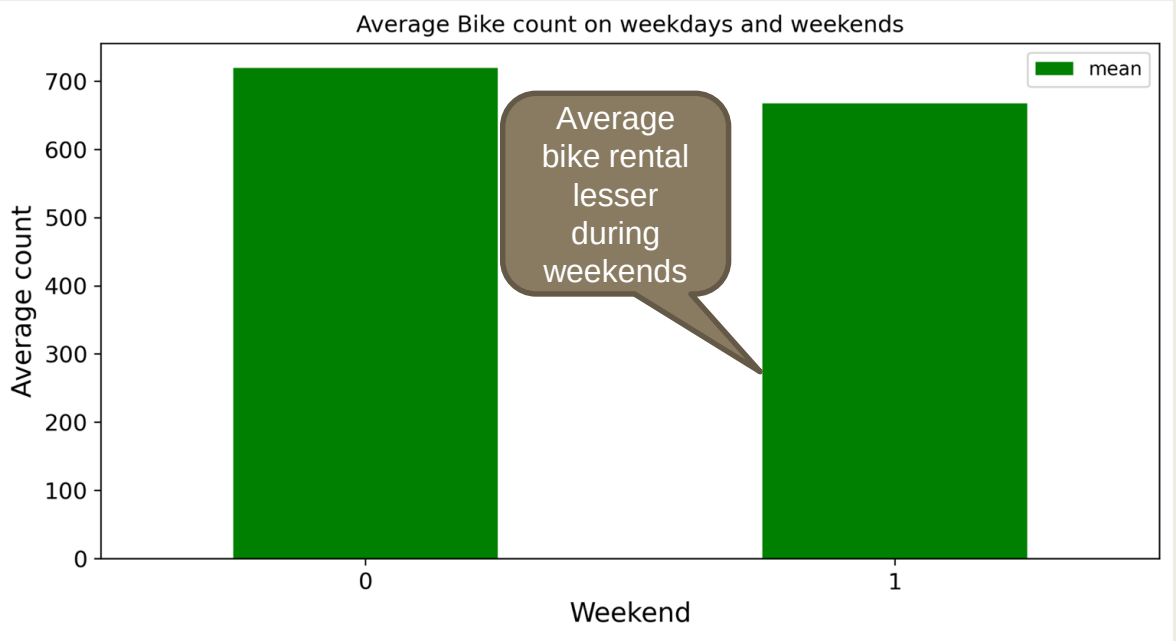


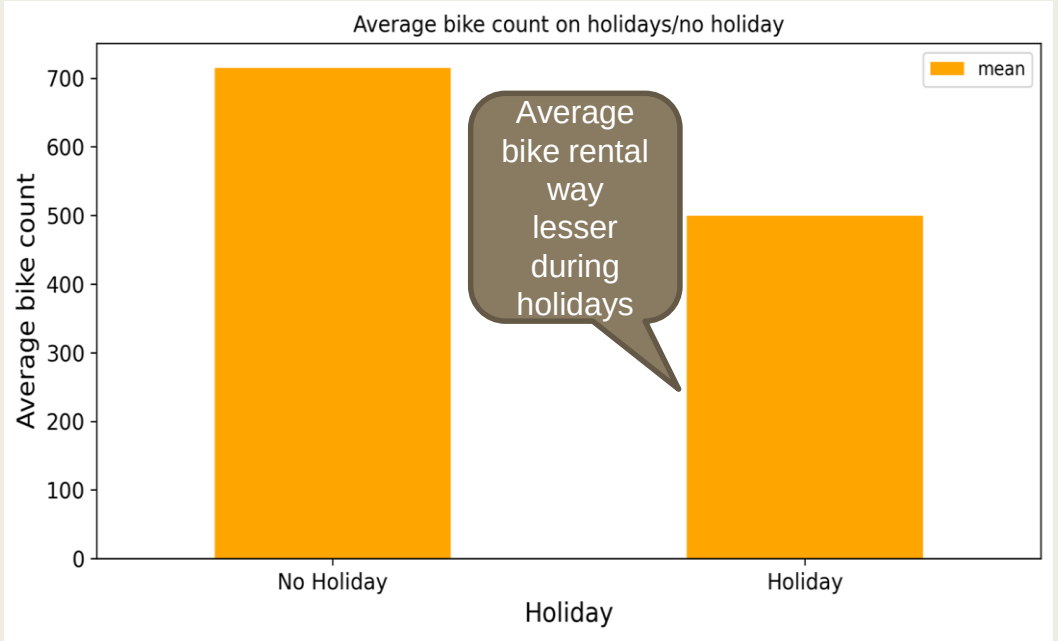


Exploratory data analysis

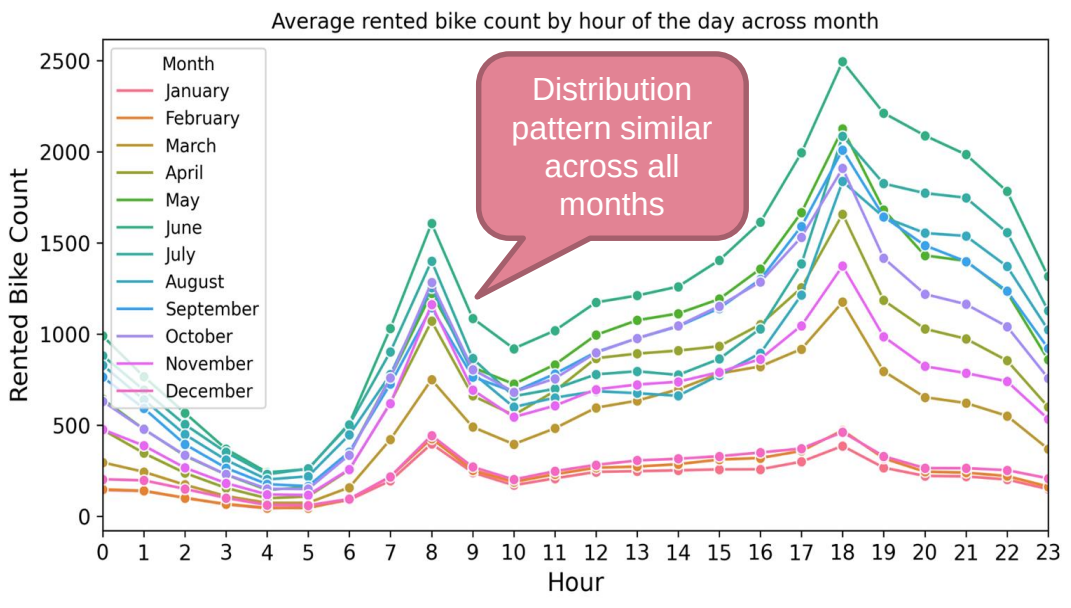


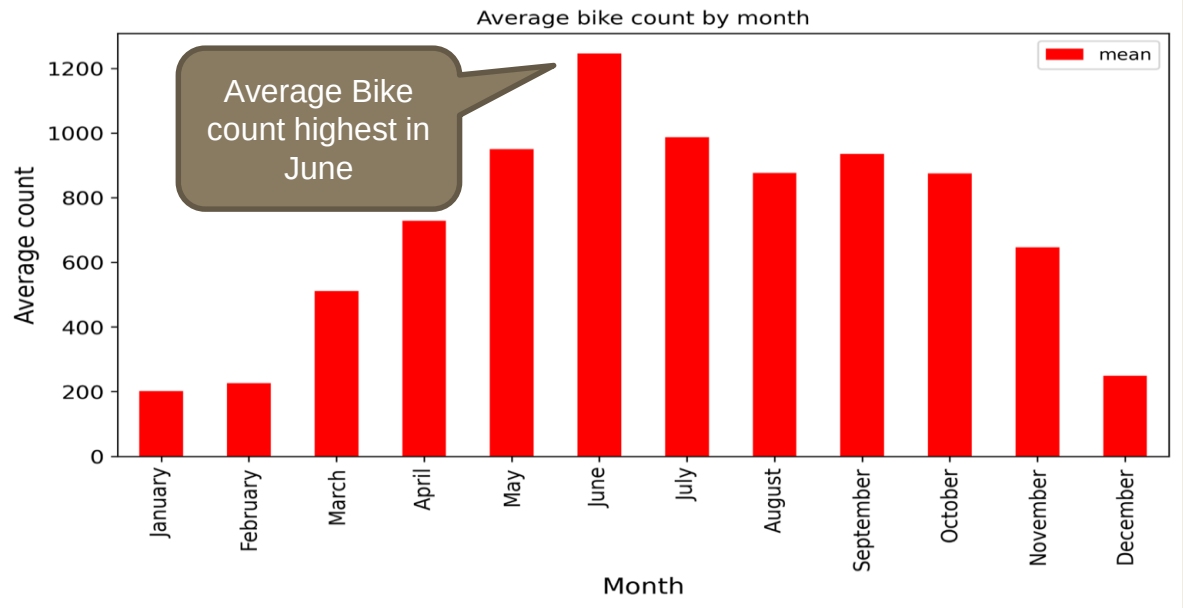


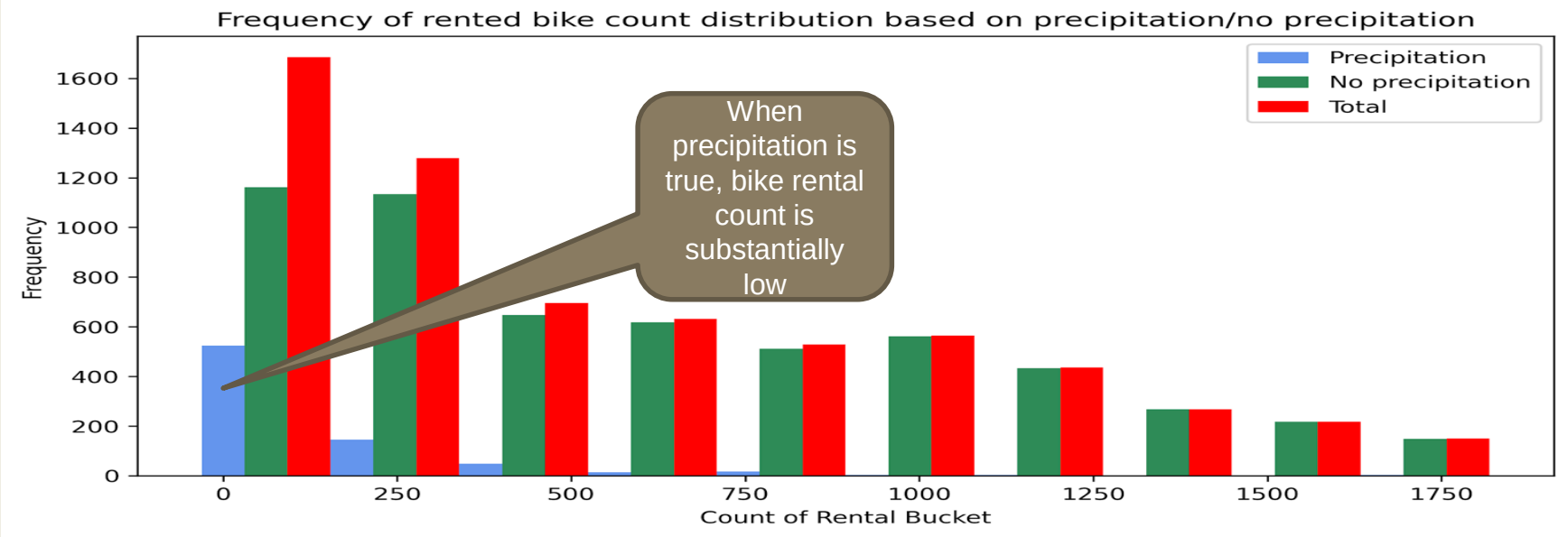




Exploratory data analysis



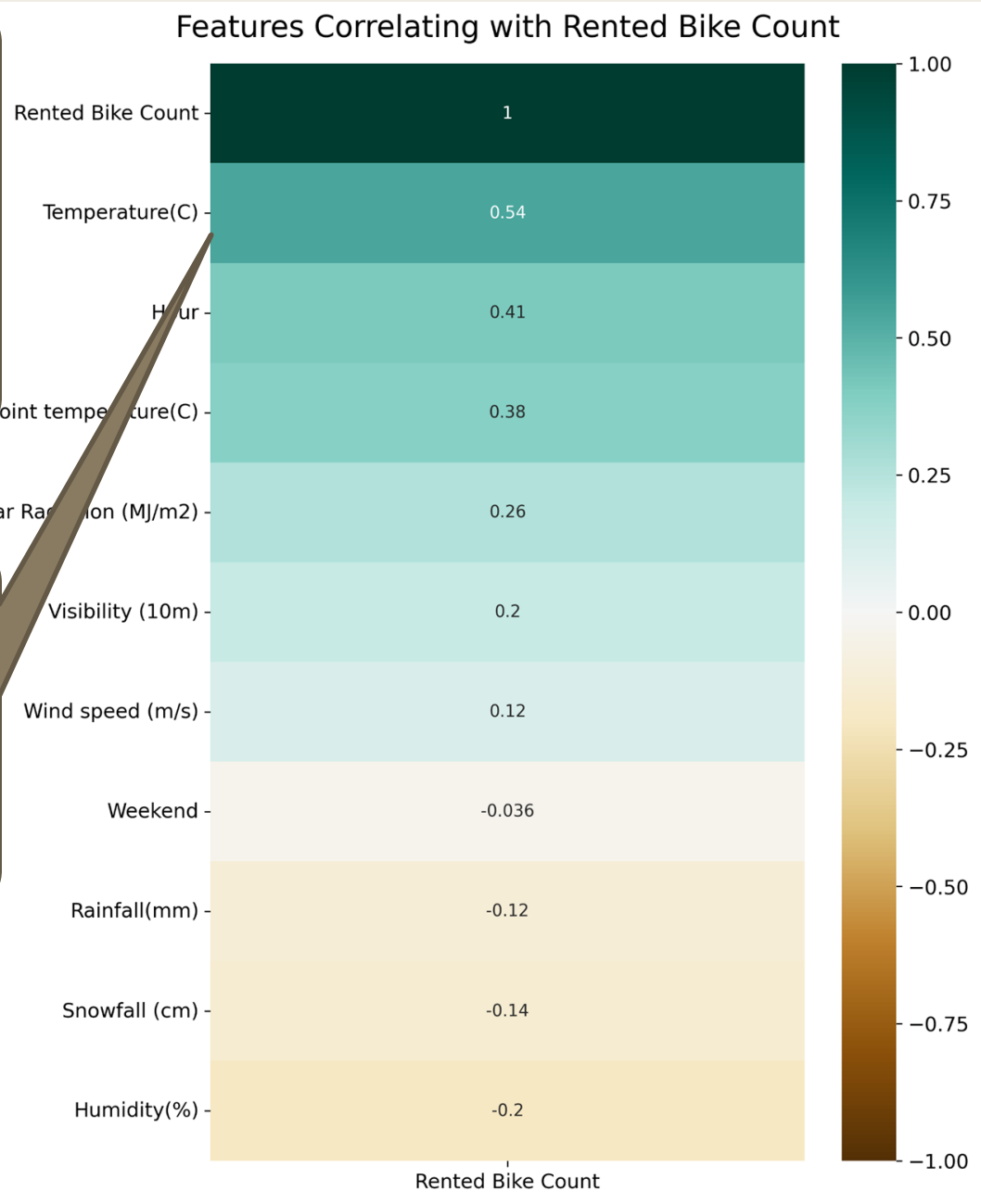




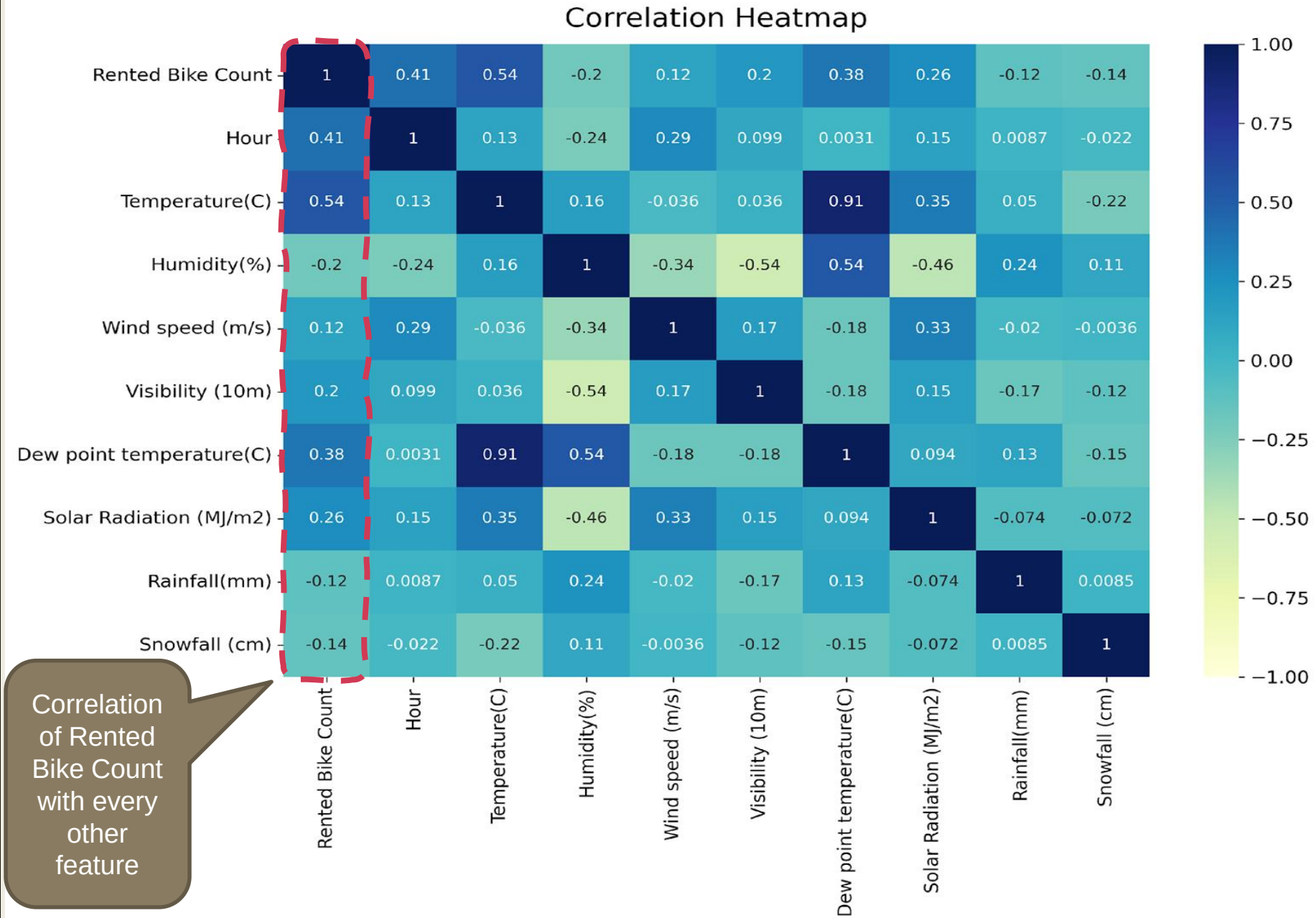
Exploratory data analysis

Dew point temperature, solar radiation positively and highly correlated, snowfall is negatively correlated to temperature

Temperature, Hour highly correlate to Rented bike count



Exploratory data analysis



Correlation of Rented Bike Count with every other feature

Correlation heatmap gives us hints on how data is distributed

Each feature correlates to each other in some way,

Finding the feature which are strong predictors

How can we distribute data so as to get more better accuracy.

Sample distribution into testing and training set

- Temperature, Dew point temperature, Hour, Seasons based buckets created for checking whether stratification helps?

Temperature:

	Overall	Stratified	Random	Rand. %error	Strat. %error
1	0.166828	0.166768	0.168584	1.052250	-0.036284
2	0.244189	0.244249	0.241374	-1.152702	0.024789
3	0.256901	0.256961	0.259231	0.907163	0.023563
4	0.273002	0.273002	0.270430	-0.942350	0.000000
5	0.059080	0.059019	0.060381	2.202869	-0.102459

Hour:

	Overall	Stratified	Random	Rand. %error	Strat. %error
1	0.291667	0.291667	0.29024	-0.489237	0.0
2	0.250000	0.250000	0.25214	0.856164	0.0
3	0.250000	0.250000	0.25000	0.000000	0.0
4	0.208333	0.208333	0.20762	-0.342466	0.0

Dew point temperature:

	Overall	Stratified	Random	Rand. %error	Strat. %error
1	0.173402	0.173373	0.170805	-1.497696	-0.016458
2	0.191667	0.191638	0.192066	0.208457	-0.014890
3	0.266553	0.266553	0.272546	2.248394	0.000000
4	0.237329	0.237300	0.234304	-1.274651	-0.012025
5	0.131050	0.131136	0.130280	-0.587979	0.065331

Seasons:

	Overall	Stratified	Random	Rand. %error	Strat. %error
Winter	0.246575	0.246575	0.244150	-0.983796	0.000000
Summer	0.252055	0.252140	0.250285	-0.701993	0.033967
Spring	0.252055	0.251998	0.255422	1.336051	-0.022645
Autumn	0.249315	0.249287	0.250143	0.331960	-0.011447

Feature Engineering

OneHotEncoding

- Employed on categorical columns like functioning day, holiday

Standard Scalar

- Scaling employed on training data so that data is scaled to same scale

Cyclic Trigonometry

- Applied on hour, seasons, day of week, month

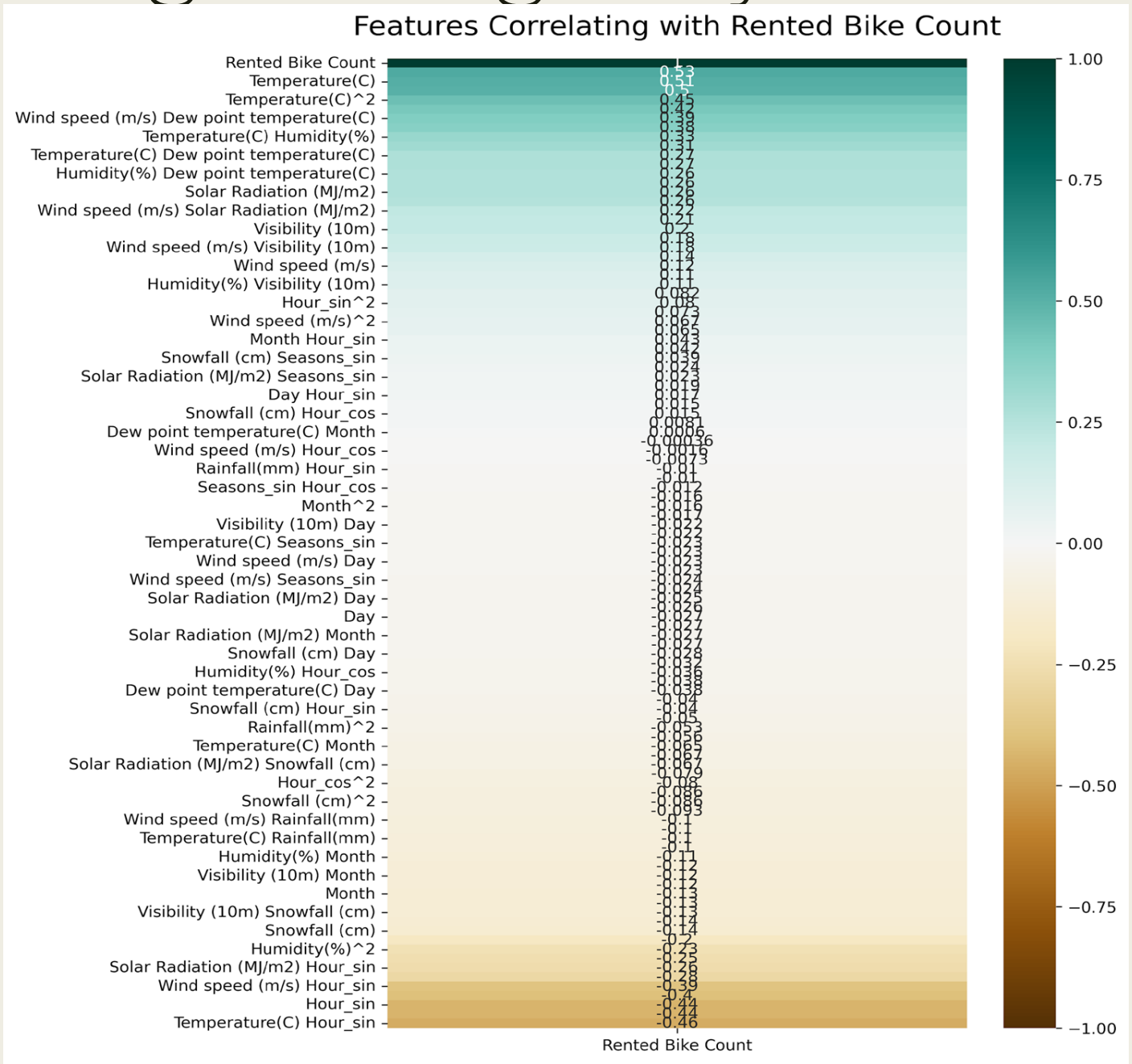
Strong predictors

- Strong predictors Identification for Imputation and prediction

Polynomial regression

- Polynomial regression employed to generate combinational features, below found to be most correlated
  - Temperature(C) & Visibility (10m) 0.529875
  - Temperature(C) & Wind speed (m/s) 0.501269
  - Temperature(C) & Hour_sin 0.461674

Feature Engineering-Polynomial Regression



Temperature Imputation

Overview

- 500 NULL Values
- Highly correlated with Dew point temperature
- Imputation methods tried
  - *Simple Imputer with mean and median*
  - *KNNImputer (Different Neighbor Size)*
  - *Bfill and Ffill (pandas transformation)*
  - *Linear Regression (Various parameter combination - Weather, Time, etc)*
  - *Kneighbor Regressor(Various parameter combination - Weather, Time, etc)*
- Linear Regression and KN models achieved up to **99.8% R2** value

Conclusion

- Simple Imputer with mean and bfill/ffill are the reference imputation techniques.
- Linear Regression and KN Regressor models with *Opt 6 to Opt 9* (marked as **green** in Table-1) yields good accuracy.
- The **Best Method** will be used to test the accuracy of the Bike Count ML model.

Temperature Imputation Models

Finding the best model

ID	Parameters	Linear Regression		KNR	
		RMSE	R2	RMSE	R2
Opt 1	Hour	11.82	0.01	12.35	-0.09
Opt 2	Hour (Sin, Cos)	11.44	0.046	12.71	-0.16
Opt 3	Hour(Sin, Cos), Month (sin)	11.67	0.052	**10.62**	0.188
Opt 4	Dew point temperature	4.83	**0.831**	5.21	0.804
Opt 5	Dew point temperature, Solar Radiation	3.56	0.914	3.94	0.888
Opt 6	Dew point temperature, Solar Radiation Humidity	1.47	0.984	0.46	0.998
Opt 7	Dew point temperature, Solar Radiation Humidity, Snowfall, Rainfall, Hour	1.37	0.987	0.65	0.996
Opt 8	Dew point temperature, Solar Radiation Humidity, Snowfall, Rainfall, Hour (Sin, Cos)	1.26	0.989	0.45	0.998
Opt 9	Dew point, Solar, Humidity, Snowfall, Rainfall, Hour(Sin, Cos), Month (sin)	1.18	0.989	0.47	0.998

BEST MODEL

Table-1: Parameters Combinations and Accuracy of Linear Regression and KN Model used for Imputation

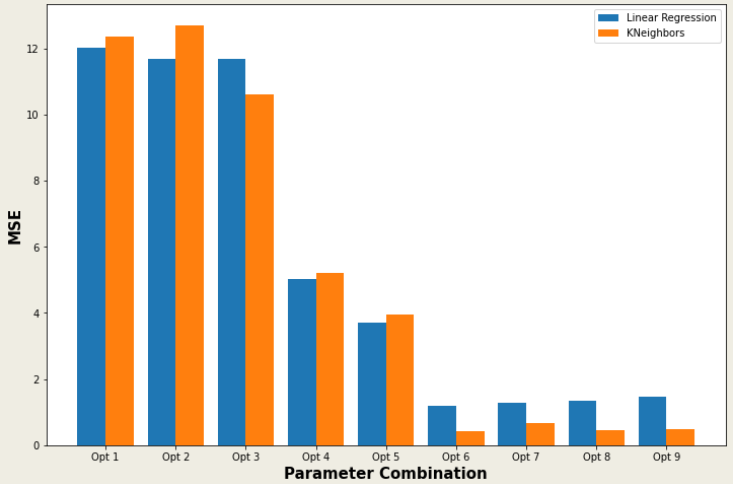


Figure-2: Root Mean Square Error Comparison for Linear Regression and KN Regressor

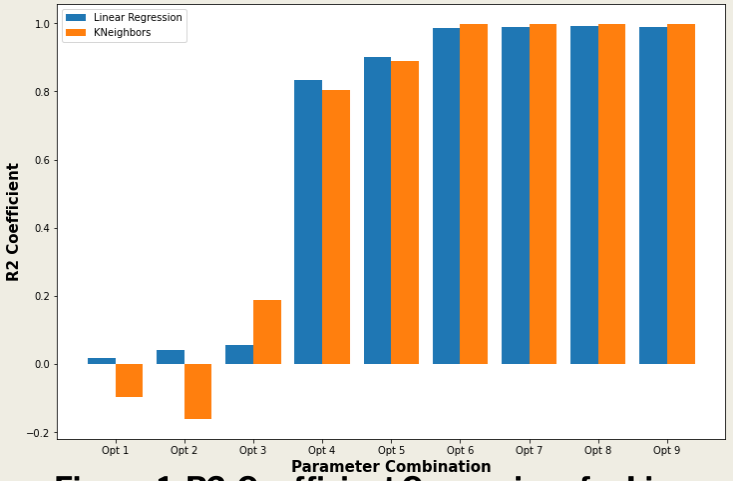


Figure-1: R2-Coefficient Comparison for Linear Regression and KN Regressor

Evaluation of Imputation Models

With Rented Bike Count ML Model (Linear Regression)

Temp Imputation Method	Rented Bike ML Model	Model based on Season	Feature Handling	Result		Remark
				RMSE	R2	
Simple Imputer	Linear Regression	One for all	Standard Scaler OHE	331.4	0.743	Mean
Simple Imputer				329.3	0.750	Median
KNNImputer				324.97	0.749	N=6
Linear Regression				326.99 315.89 323.68 325.71	0.733 0.749 0.736 0.745	Opt 6 Opt 7 Opt 8 Opt 9
KNeighbor Regressor				326.23 326.50 329.85 327.26	0.758 0.751 0.748 0.755	Opt 6 Opt 7 Opt 8 Opt 9
bfill				316.4	0.743	
ffill				330.2	0.740	

Strategy for ML Model

- Test with Various Models
  - *Linear Regression, RidgeCV, Lasso, ElasticNet*
  - *SGDRegressor*
  - *KNRegressor*
  - *XGBoost (as Reference)*
- Split the data and use separate models
  - *Split with Seasons (Same model or Separate model for each season)*
  - *Split with Precipitation and no-Precipitation*
  - *Split with Day and Night Time, Offtime/Peaktime, Weekdays/Weekend*
  - *Holiday/No Holiday*
- Perform K-Fold Cross-Validation (10 folds)
- Test against the best Temperature Imputation Method - **KNR with Opt8**
- Perform Regularization using Lasso, RidgeCV
- Hyperparameter Tuning with GridSearchCV (SGD, ElasticNet)
- Check whether Stratified Sampling has any effect on the model
- Compare the metric for choosing the best model

Model for Rented Bike Count Performance Evaluation

No Split

Temp Imputation Method	Rented Bike ML Model	Model Split	Feature Handling	Validation Set Result	
				RMSE	R2
KNR (Opt8)	Linear Regression	One for all	Cyclic Feature Std Scaler, OHE	418.97	0.546
KNR (Opt8)	**KNR**	**One for all**	**Cyclic Feature Std Scaler, OHE**	250.23	0.843
KNR (Opt8)	RidgeCV	One for all	Cyclic Feature Std Scaler, OHE	418.73	0.546
KNR (Opt8)	Lasso	One for all	Cyclic Feature Std Scaler, OHE	418.75	0.544
KNR (Opt8)	ElasticNet	One for all	Cyclic Feature Std Scaler, OHE	452.0	0.462
KNR (Opt8)	SGDRegressor	One for all	Cyclic Feature Std Scaler, OHE	419.04	0.547

Final Model on Testset: KNeighborsRegressor,
RMSE = 289.80, R2 =0.814

No Split - Stratified Sampling (Hour)

- Stratified Sampling on Hour - Bins of 6 Hours

Temp Imputation Method	Rented Bike ML Model	Model Split	Feature Handling	Validation Set Result	
				RMSE	R2
KNR (Opt8)	Linear Regression	One for all	Cyclic Feature Std Scaler, OHE	404.55	0.584
KNR (Opt8)	**KNR**	**One for all**	**Cyclic Feature Std Scaler, OHE**	**266.59**	**0.829**
KNR (Opt8)	RidgeCV	One for all	Cyclic Feature Std Scaler, OHE	404.6	0.583
KNR (Opt8)	Lasso	One for all	Cyclic Feature Std Scaler, OHE	404.7	0.583
KNR (Opt8)	ElasticNet	One for all	Cyclic Feature Std Scaler, OHE	446.03	0.494
KNR (Opt8)	SGDRegressor	One for all	Cyclic Feature Std Scaler, OHE	404.78	0.583

Final Model on Testset : KNeighborsRegressor,
RMSE = 300.82, R2 =0.805

Models With Data Splits Performance Evaluation

Season-based

Rented Bike ML Model	Model Split	Feature Handling	Results of CV		Remark
			RMSE	R2	
Linear Regression	Separate for Seasons	Cyclic Feature Std Scaler, OHE	441.8 346.79 365.74 93.59	0.539 0.698 0.677 0.509	Summer Spring Autumn Winter
KNR	**Separate for Seasons**	**Cyclic Feature Std Scaler, OHE**	**293.29 220.33 256.51 73.1**	**0.796 0.878 0.836 0.696**	**Summer Spring Autumn Winter**
SGDRegressor	Separate for Seasons	Cyclic Feature Std Scaler, OHE	442.37 346.83 365.88 93.7	0.538 0.698 0.677 0.508	Summer Spring Autumn Winter

Final Model on Testset: KNeighborsRegressor
RMSE = 292.16, R2 =0.811

Precipitation/No Precipitation

Rented Bike ML Model	Model Split	Feature Handling	Result of CV		Remark
			Min RMSE	Max R2	
Linear Regression	Separate for Precipitation and No Precipitation	Cyclic Feature for Hour, Date Std Scaler for Numerical cols OHE for Categorical cols	143.31 423.33	0.351 0.551	Precipitation No Precipitation
KNR			**117.85** **257.61**	**0.55** **0.839**	Precipitation No Precipitation
RidgeCV			143.03 423.18	0.344 0.552	Precipitation No Precipitation
Lasso			143.31 422.89	0.326 0.551	Precipitation No Precipitation
ElasticNet			149.26 460.48	0.288 0.458	Precipitation No Precipitation
SGDRegressor			146.02	0.312 0.552	Precipitation No Precipitation

Final Model on Testset: KNeighborsRegressor
RMSE = 277.06, R2 =0.83

Weekday/Weekend

Temp Imputation Method	Rented Bike ML Model	Model Split	Feature Handling	Result of CV		Remark
				Min RMSE	Max R2	
KNR (Opt8)	Linear Regression	Separate for Weekdays and Weekends	Cyclic Feature for Hour, Date Std Scaler for Numerical cols OHE for Categorical cols	375.93 436.21	0.643 0.53	Weekdays Weekends
KNR (Opt8)	**KNR**			**228.79** **264.84**	**0.862** **0.844**	Weekdays Weekends
KNR (Opt8)	RidgeCV			376.28 436.21	0.643 0.531	Weekdays Weekends
KNR (Opt8)	Lasso			375.55 435.89	0.641 0.528	Weekdays Weekends
KNR (Opt8)	ElasticNet			393.84 461.34	0.548 0.439	Weekdays Weekends
KNR (Opt8)	SGDRegressor			375.15 437.27	0.644 0.526	Weekdays Weekends

Final Model on Testset: KNeighborsRegressor
RMSE = 294.12, R2 =0.808

Day/Night

Temp Imputation Method	Rented Bike ML Model	Model Split	Feature Handling	Result of CV		Remark
				Min RMSE	Max R2	
KNR (Opt8)	Linear Regression	Separate for Day and Night	Cyclic Feature for Hour, Date Std Scaler for Numerical cols OHE for Categorical cols	403.81 310.03	0.584 0.743	Day Night
KNR (Opt8)	**KNR**			**298.53 199.86**	**0.782 0.886**	**Day Night**
KNR (Opt8)	RidgeCV			403.73 310.26	0.584 0.744	Day Night
KNR (Opt8)	Lasso			403.54 310.01	0.582 0.743	Day Night
KNR (Opt8)	ElasticNet			436.56 354.81	0.455 0.665	Day Night
KNR (Opt8)	SGDRegressor			401.4 310.53	0.576 0.740	Day Night

Final Model on Testset: KNeighborsRegressor
RMSE = 293.34, R2 =0.809

Time-based

Split By (12 AM to 6 AM - **OFFTIME**) , (7 AM to 11 PM - **PEAKTIME**)

Temp Imputation Method	Rented Bike ML Model	Model Split	Feature Handling	Result of CV		Remark
				Min RMSE	Max R2	
KNR (Opt8)	Linear Regression	Separate for Peak and Off time. During 12 AM to 6 AM	Cyclic Feature for Hour, Date Std Scaler for Numerical cols OHE for Categorical cols	426.34 141.55	0.566 0.715	Peaktime Offtime
KNR (Opt8)	**KNR**			**292.13** **112.84**	**0.813** **0.819**	Peaktime Offtime
KNR (Opt8)	RidgeCV			428.25 141.55	0.567 0.715	Peaktime Offtime
KNR (Opt8)	Lasso			427.6 142.28	0.568 0.712	Peaktime Offtime
KNR (Opt8)	ElasticNet			476.88 162.99	0.467 0.576	Peaktime Offtime
KNR (Opt8)	SGDRegressor			427.53 141.89	0.568 0.713	Peaktime Offtime

Final Model on Testset: Lasso
RMSE = 199.73, R2 =0.722

Final Model on Testset: KNeighborsRegressor
RMSE = 176.34, R2 =0.864

Holiday/No Holiday

Temp Imputation Method	Rented Bike ML Model	Model Split	Feature Handling	Result of CV		Remark
				Min RMSE	Max R2	
KNR (Opt8)	Linear Regression	Separate for Holiday and Non-Holiday	Cyclic Feature for Hour, Date Std Scaler for Numerical cols OHE for Categorical cols	**251.70** 420.19	**0.73** 0.543	Holiday No Holiday
KNR (Opt8)	**KNN**			105.96 **258.22**	**0.949** **0.839**	Holiday No Holiday
KNR (Opt8)	RidgeCV			**248.85** 420.19	**0.714** 0.542	Holiday No Holiday
KNR (Opt8)	Lasso			**250.65** 419.91	**0.725** 0.542	Holiday No Holiday
KNR (Opt8)	ElasticNet			**287.18** 453.38	0.61 0.465	Holiday No Holiday
KNR (Opt8)	SGDRegressor			252.02 420.16	0.713 0.541	Holiday No Holiday

Final Model on Testset: KNeighborsRegressor
RMSE = 288.79, R2 =0.815

Final ML Model

Rented Bike ML Model	Linear Regression Performance (LR, RidgeCV, Lasso, ENet)	Best Model
No Split	**Lasso RMSE = 474.86, R2 =0.502**	**KNR RMSE = 289.80, R2 =0.814**
No Split, Stratified	**LR RMSE = 462.81, R2 =0.539**	**KNR RMSE = 300.82, R2 =0.805**
Season-based	**RidgeCV, LR, Lasso, RidgeCV RMSE = 411.98, R2 =0.624**	**KNR RMSE = 289.80, R2 =0.814**
Precipitation/No Precipitation	**RidgeCV/Lasso RMSE=455.15, R2=0.542**	**KNR RMSE = 277.06, R2 =0.83**
Weekdays/ Weekend	**Lasso/Lasso RMSE=475.08, R2=0.501**	**KNR RMSE = 294.12, R2 =0.808**
Day/Night	**Lasso/LR RMSE=419.98, R2=0.610**	**KNR RMSE = 293.34, R2 =0.809**
Time-based	**Lasso/LR RMSE=426.45, R2=0.598**	**KNR RMSE = 287.07, R2 =0.817**
Holiday/No Holiday	**RidgeCV/Lasso RMSE=471.48, R2=0.508**	**KNR RMSE = 288.79, R2 =0.815**


Thank you