Image Credit: Tiia Monto[1]

DSTA Coursework 2: London Cycles Dataset Principal Component analysis

Montel Moore | MSc Data Science | 13/03/2021

PHASE 1

1.1. RECAP: DATASET DESCRIPTION AND FEATURE IMPORTANCE

Previous analysis of the dataset concluded that the most important numerical dimensions were: humidity, temperature and hour of day. The Figure 1 is a scatter graphs displaying the relationships between independent variables hour, temperature, humidity and the chosen depended variable, number of bike shares per hour. It can been seen that there is a multimodal relationship between hour and bike shares, and weak linear relationships between temperature and humidity percentage.

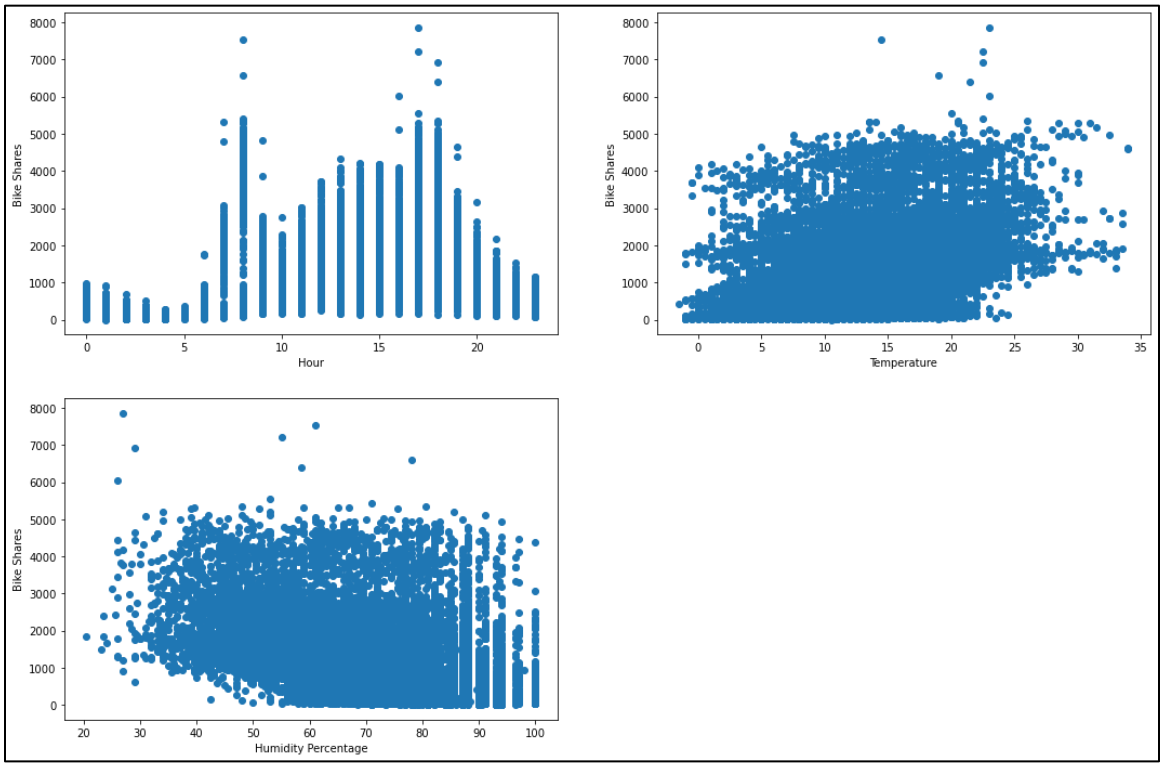

Figure 1: Scatter plots between bike shares and the three variables.

The distribution of the variables are shown in Figure 2. Hour has an even distribution between 0 and 24, while temperature has a positive skew (mean = 12.5 , median = 12.5, mode = 13.0, range = -1.5 to 34.0) and humidity percentage has a negative skew (mean = 72.5 , median = 74.5, mode = 88.0 , range = 20.5 to 100.0)

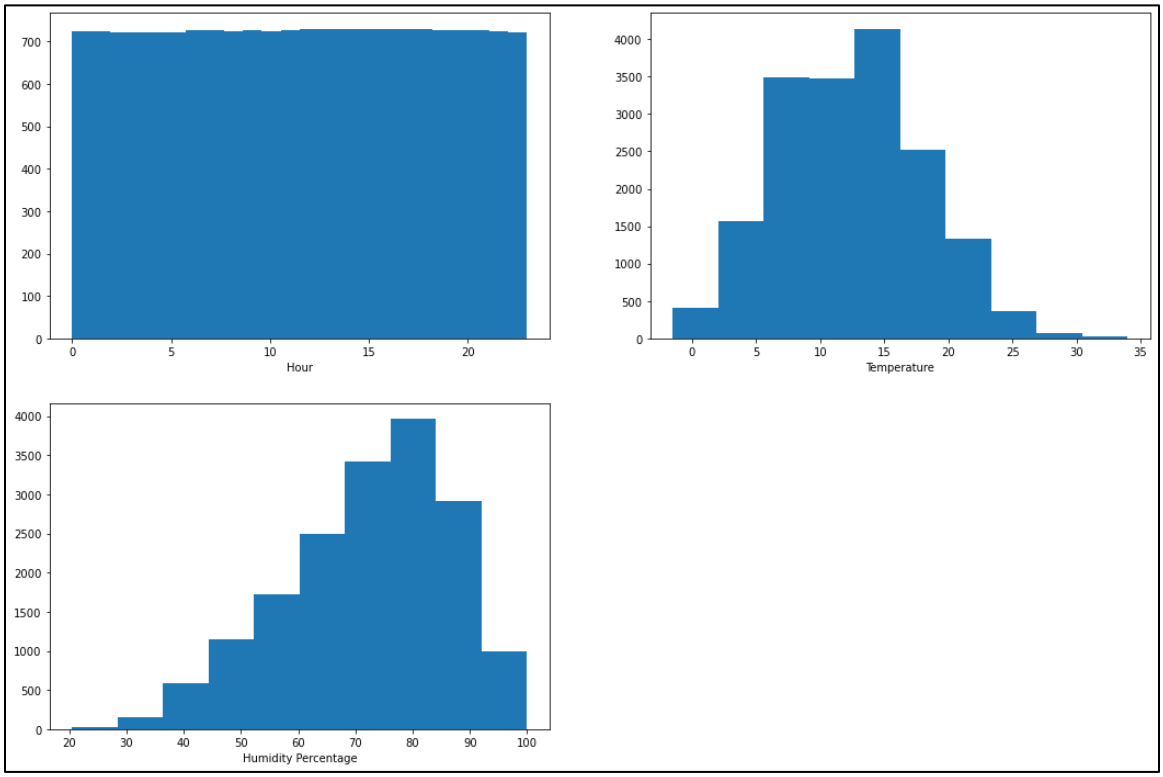

Figure 2: Histograms for the independent variables

Combined, these three variables should provide a good dimensionality reduction, however there are more variables that impact the number of bike shares, such as day of week, rainfall and time of year.

Phase 2

2.1. METHODOLOGY

Three methods were used to perform dimensionality reduction: Linear PCA, Kernel PCA and FAMD (factor analysis of mixed data). Kernel PCA and FAMD were used as a matter of interest; kernel PCA allows us to better linearize non-linear patterns of the data, while the goal of FAMD is to linearize categorical variables in combination with numerical variables. The FAMD reduction will be compared with PCA performed on one-hot encoded data. The technical detail of FAMD and kernel PCA are beyond the scope of this report.

In order to assess the effectiveness of the dimensionality reduction, models were initially tested on the following data sets:

1. Reduced dataset (Independent variables: Hour, Temperature, Humidity)
2. Continuous Numeric variables only (Independent variables: Hour, Temperature, perceived temperature precipitation, wind speed, humidity)
3. All variables, with categorical variables one-hot encoded

4. All variables, with categorical variables not one-hot encoded (accepted input for the FAMD function)

Kernel-PCA was used to transform each dataset (except for dataset 4, which was transformed with FAMD), combined with an extremely randomized trees regressor model to predict the number of bike shares per hour. Linear and ridge regression models were also used but showed inferior results to the extra-trees model, and were removed from the code appendix to save space. The optimal parameters for PCA were determined for each model by using a 2-fold grid search. Linear, polynomial (2nd and 3rd degree), cosine kernels and the number of principal components to feed the model (1, 2, or 3) were tested. Additionally each dataset (except for 4 due to type errors) was fed "raw" into the extra-trees model, without the dimensionality reduction applied.

```
pca = KernelPCA()
extra_trees = ExtraTreesRegressor()
famd = FAMD()

pca_params = [
    {'pca__n_components': [1,2,3], 'pca__kernel': ["linear"]},
    {'pca__n_components': [1,2,3], 'pca__kernel': ["poly"], 'pca__degree': [2,3]},
    {'pca__n_components': [1,2,3], 'pca__kernel': ["cosine"]}]

famd_params = {'famd__n_components': [1,2,3]}

pca_pipe = Pipeline(steps = [["scale", MinMaxScaler()],['pca', pca], ['lr', extra_trees]])

pca_prescaled_pipe = Pipeline(steps = [['pca', pca], ['lr', extra_trees]])

famd_pipe = Pipeline(steps = [['famd', famd], ['lr', extra_trees]])

bare_bones_grid = GridSearchCV(extra_trees, param_grid = {},
                               scoring = ("neg_root_mean_squared_error"), cv = KFold(2, shuffle = True))

pca_grid = GridSearchCV(pca_pipe, pca_params,
                        scoring = ("neg_root_mean_squared_error"), cv = KFold(2, shuffle = True))

pca_prescaled_grid = GridSearchCV(pca_prescaled_pipe, pca_params,
                                  scoring = ("neg_root_mean_squared_error"), cv = KFold(2, shuffle = True))

famd_grid = GridSearchCV(famd_pipe, famd_params,
                         scoring = ("neg_root_mean_squared_error"), cv = KFold(2, shuffle = True))
```

Figure 3: Grid search and pipeline code block for KernelPCA/FAMD, followed by extra trees regressor. The extra trees regressor was used on both PCA/FAMD reduced and non-reduced data.

The dim-reduced models were compared with the "raw" models with the ratio: **(RMSE/mean(bike shares))** and the "explained variance" metrics. The **RMSE/mean(bike shares)** ratio shows the average deviation from the mean i.e. it is a measure of how proportionately bad predicted y is from estimated y. This measure is important particularly in 3.2, where the dataset is divided by hour, and each "hour" has a difference in mean bike shares. This is an experimental metric, therefore the main focus will be the more established explained variance metric.

All variables were scaled to size 0-1 prior to dimensionality reduction, due to the difference in scale between the variables.

Phase 3

3.1. Initial Results

In general, the non-dimensionality-reduced models were better predictors of bike shares. The full data set with one-hot encoded variables performed significantly well, with an explained variance score of 0.95. This is a significantly better score than the three-variable dataset with PCA, which has an explained variance score of 0.61. The three variable dataset was tested with and without PCA, and the former had a slightly higher explained variance (0.61 vs 0.59).

The experimental FAMD-reduced model displayed the worst performance, and the one-hot encoded dataset reduced by PCA displayed the 2nd worst performance. One explanation for this is the amount of information held by these datasets is much larger than the three-variable/numeric variable only dataset, while the number of principal components was limited to three. Due to the relatively higher number of dimensions, it is highly likely that these models would have shown stronger performance if more principal components were permitted.

Data Range	Model	RMSE-mean bike shares ratio	Explained variance	Rank
All hours	Hour, Temp, Humidity PCA and extra trees regressor	0.59	0.61	4
	All numeric variables with PCA and extra trees regressor	0.57	0.65	3
	All variables (one-hot categories) with PCA and extra trees regressor	0.75	0.37	6
	Numerical and non-numerical variables FAMD and extra trees regressor	0.81	0.26	7
	Three variables without PCA , but with extra trees regressor	0.61	0.59	5
	All numeric variables without PCA, but with extra trees regressor	0.54	0.68	2
	Binarized variables without PCA, but with and extra trees regressor	0.21	0.95	1
Hours 8-19	Hour, Temp, Humidity PCA and extra trees regressor	0.51	0.28	2
	All variables (one-hot categories) w/o PCA, but with extra trees regressor	0.19	0.91	1

Table 1: Results for each model.

Figure 4 shows the graphical relationship between bike shares and each principal component for the main reduced dataset (rank 4). It is obvious to see that there is only a weak relationship between the number of bike shares and each principal component. One reason for this may be the "spread" in the number of bike shares per hour, as shown in Figure 5. Though this proves that "hour" is an important variable for predicting bike shares, it tells us that there may be other variables that influence the number of bike shares, and that some "hours" are more predictable than others due to the number of outliers. This hypothesis was initially tested on a dataset reduced to hours 8-19, as shown in Table 1. This dataset displayed a lower explained variance, but stronger RMSE-mean bike shares ratio. Further analysis of the hour variable is shown in 3.2.

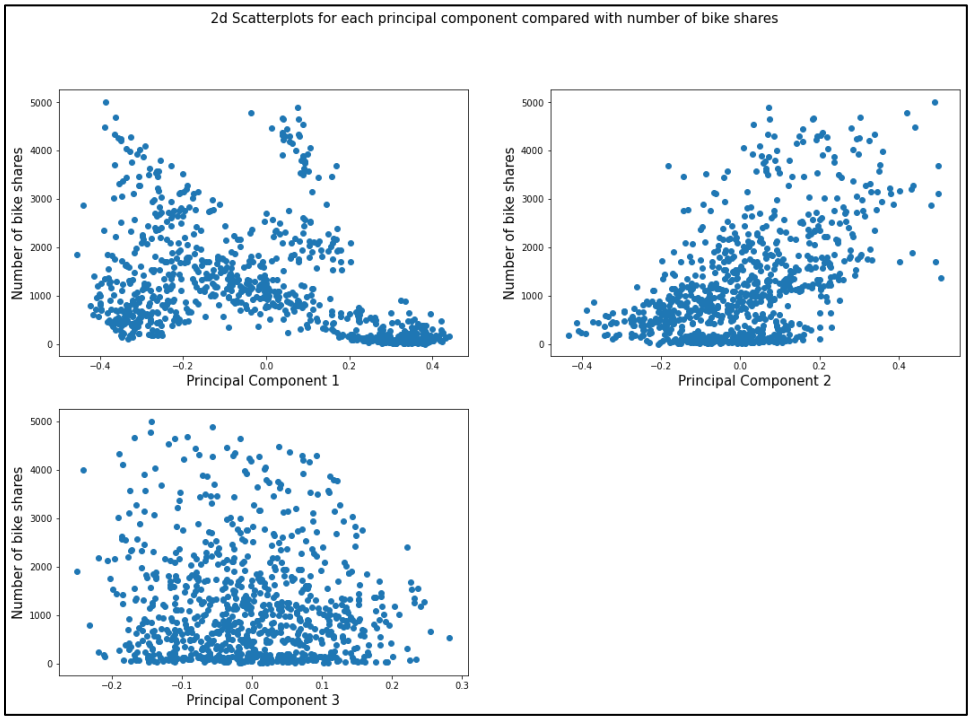

Figure 4: Bike shares/each principal component

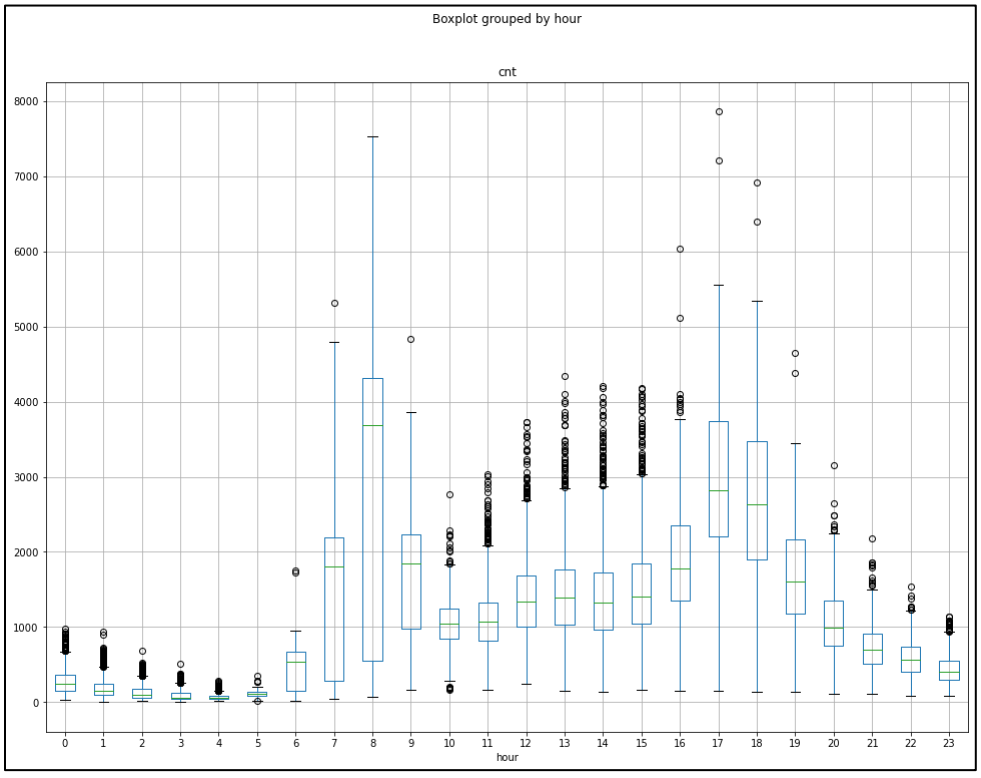

Figure 5: distribution of bike shares for each hour of the day.

3.2. Further Analysis

Out of interest, the dataset was split by hour, and the PCA + regression was performed on is weekend, temperature and humidity. The results of each model are shown in Table 2.

Data Range	RMSE-mean bike shares ratio		Explained variance	
	is_weekend, temp, humidity PCA + extra trees regressor	All variables + extra trees regressor	is_weekend, temp, humidity PCA + extra trees regressor	All variables + extra trees regressor
Hour 0	0.38	0.30	0.54	0.74
Hour 1	0.48	0.37	0.63	0.76
Hour 2	0.49	0.38	0.68	0.80
Hour 3	0.49	0.38	0.63	0.80
Hour 4	0.39	0.34	0.62	0.73
Hour 5	0.28	0.23	0.11	0.53
Hour 6	0.34	0.22	0.63	0.84
Hour 7	0.37	0.20	0.64	0.87
Hour 8	0.36	0.20	0.70	0.90
Hour 9	0.26	0.16	0.59	0.81
Hour 10	0.24	0.18	0.46	0.70
Hour 11	0.28	0.19	0.58	0.76
Hour 12	0.25	0.22	0.63	0.76
Hour 13	0.27	0.23	0.70	0.80
Hour 14	0.28	0.22	0.69	0.81
Hour 15	0.28	0.22	0.66	0.78
Hour 16	0.24	0.19	0.61	0.77
Hour 17	0.22	0.17	0.68	0.82
Hour 18	0.25	0.17	0.72	0.86
Hour 19	0.24	0.19	0.70	0.82
Hour 20	0.24	0.21	0.68	0.77
Hour 21	0.27	0.23	0.58	0.73
Hour 22	0.28	0.21	0.54	0.75
Hour 23	0.37	0.23	0.34	0.73
Average	**0.31**	**0.24**	**0.60**	**0.78**

Table 2: Further analysis results.

Table 2 clearly shows that the dimensionality reduction was more effective for some hours, and less effective for other hours, when compared to the non-divided dataset shown in Table 1. The addition of the binary “is_weekend” variable likely assisted in boosting the predictive power of the model, as there is a significant association with cycling on weekdays vs weekends.

Figure 6 and Figure 7 show scatterplots for bikeshares/principal components for hour 5 and hour 18 respectively. These hours were chose because they represent the highest and lowest explained variances. It is clear to see that the principal components of hour 18 contain more information related to bike shares, when compared to hour 5. This serves as proof that some hours are more predicable than others.

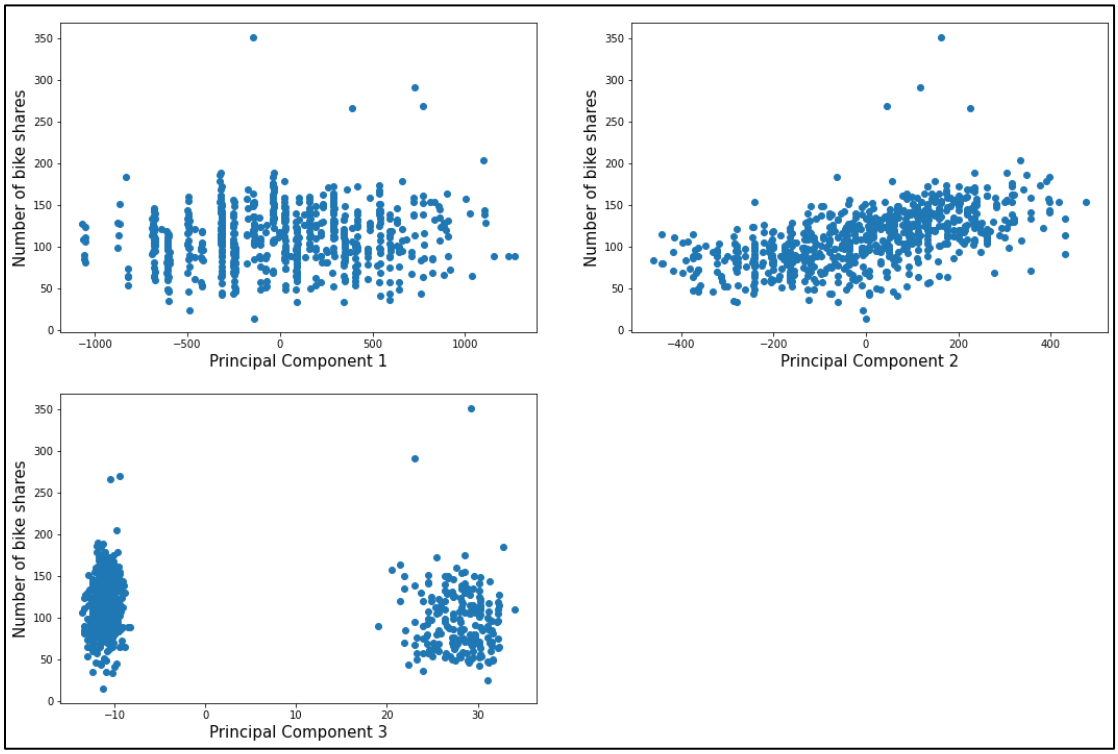

Figure 6: Scatterplots for each principal component compared with number of bike shares during hour 5

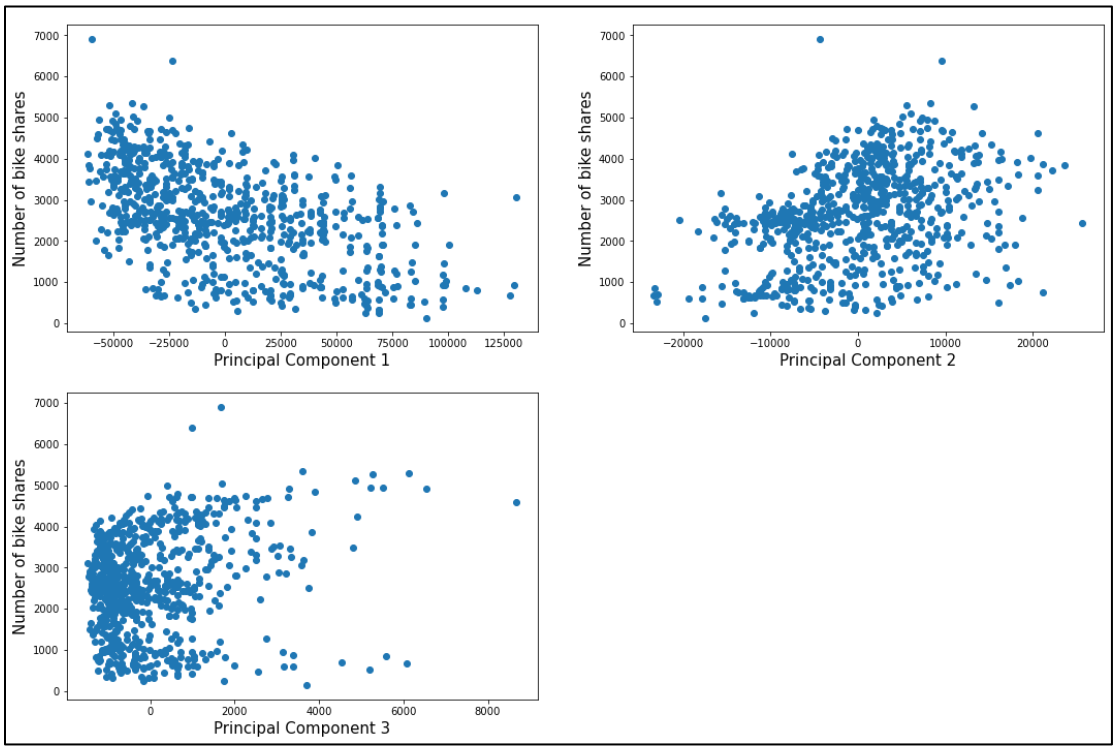

Figure 7: Scatterplots for each principal component compared with number of bike shares during hour 18

The addition of the "is_weekend" variable makes for an unfair comparison due to the additional information that it provides, therefore another model was created. This model was the original dataset (hours not divided) with four variables: hour, is weekend, temperature and humidity.

	RMSE-mean bike shares ratio		Explained Variance	
Hour Subset	hour, is_weekend, temp, humidity PCA + extra trees regressor	All variables + extra trees regressor	hour, is_weekend, temp, humidity PCA + extra trees regressor	All variables + extra trees regressor
All hours	0.35	0.21	0.86	0.95

Table 3: four-variable PCA

The four-variable dataset showed a higher explained variance compared to the "hour-divided" datasets. An explanation for this is the fact that "hour" itself is a good predictor of the number of bike shares. The four-variable PCA included "hour" in its principle components, while the "hour-divided" datasets retained the information of hour due to their nature of separation, but could not hold that information in their principle components. On average, the "hour-divided" datasets showed a lower RMSE-mean ratio, which indicates that the prediction errors were proportionally lower than the full, four-variable dataset.

3.3. CONCLUSION

Effectiveness of dimensionality reduction

The London bike-sharing dataset contains a high number of variables that correlate to the number of bike shares. Though there was some success in reducing the dataset to three variables before applying PCA, it would be sensible to include more variables, particularly those relating to time and weather.

The appendix from the previous report showed significant correlations between bike shares and other variables such as day-of-week and season. These variables were dummy encoded and performed well with extra-trees regression **without** PCA. It would be interesting to perform a further analysis to see if there exists a lower dimensional approximation for the full dataset, where a model would be able to predict bike shares accurately. In my opinion, an early hypothesis would be a dummy encoded dataset which captures the following information:

- Time of day
- Day of week
- Month of year (this contains seasonal information)
- Quality of weather temperature
- Quality of weather precipitation

There were ambitions for the added variable, precipitation, to have a positive impact on the predictive power of the model, however this was not the case. This is likely because precipitation might have more of an effect over a timespan longer than one hour. A future

analysis may require a dummy encoded variable that represents if there was rain on the **day,** rather than hour. Additionally, the precipitation was measured at one weather station in London, rather than multiple. It may be sensible to gain an average precipitation across multiple weather stations. It may also be of sensible to cluster the precipitation levels into bins (0.2mm precipitation is not noticeable compared to 2.2mm precipitation)

There is scope to transform this into a classification task, with bike shares placed into "bins" i.e. 0-1000, 1000-2000, 3000-4000 and performing logistic regression/extra trees classification. It is difficult to regress on a vast amount of data, and performing a classification task may simplify the model at expense of accuracy.

Further limitations include the kernel used for PCA, due to time constraints. It is of interest to explore kernels such as RBF or sigmoid.

REFERENCES

1. Monto, T., 2018. *Santander bicycles on the Exhibition Road in London.*. [image] Available at: <https://commons.wikimedia.org/wiki/File:Santander_Cycles.jpg> [Accessed 19 February 2021].
2. https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html
3. https://link.springer.com/content/pdf/10.1007/s10994-006-6226-1.pdf - Extremely random trees