1 Principle Component Analysis

Principal Component Analysis (PCA) is a technique used to reduce the dimensionality of a data set. In addition, PCA finds a linear projection of high dimensional data into a lower-dimensional subspace such as the variance retained is maximized and the least-square reconstruction error is minimized. Furthermore, " The PCs are essentially the linear combinations of the original variables, the weights vector in this combination is actually the eigenvector found which in turn satisfies the principle of least squares"(Property Component Analysis Tutorial).

Importance of components:						
	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.7015	1.1025	0.9219	0.7758	0.65553	0.08902
Proportion of Variance	0.4825	0.2026	0.1416	0.1003	0.07162	0.00132
Cumulative Proportion	0.4825	0.6851	0.8267	0.9271	0.99868	1.00000

Table 1: This table gives the standard deviation, proportion of variance explained by each of the principal component, and the cumulative proportion of variance explained. Also this table was generated with the use of the "prcomp" function which is used to numerically stable routine that returns a "prcomp object" that contains the square-root of the eigenvalues (sdev), the eigenvectors (rotation), and the scores (x) (More Principle Components).

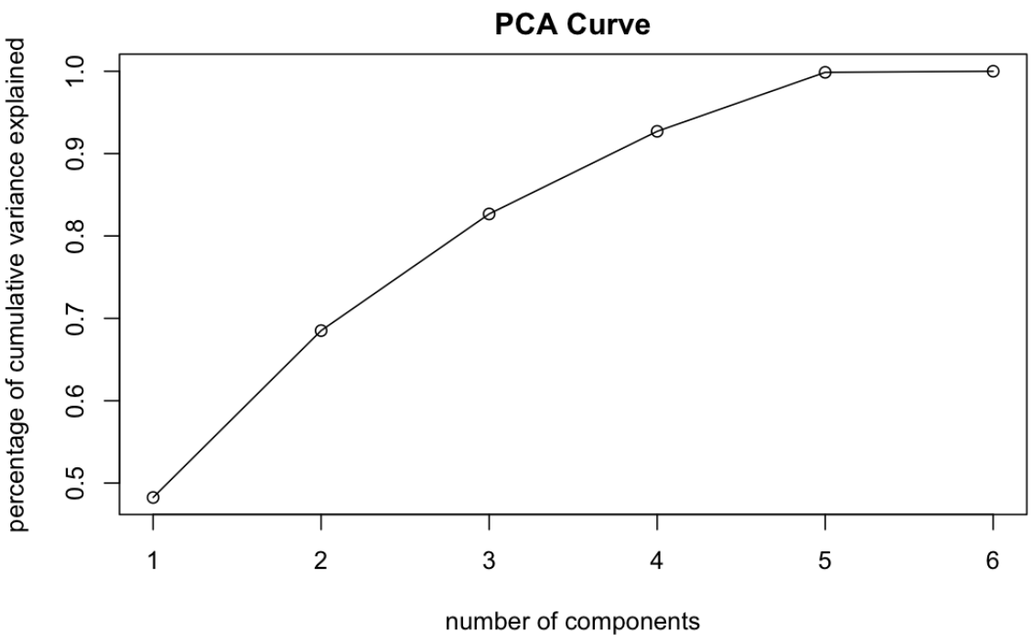

Figure 1: This figure showcases how percentage of the variance in the data explained as we add principal components. For example, out PCA1 48 percent of the data set. The PCA2 explains 20 percent and PCA3 explains 14 percent of the data. The total of the 3 components togther 82 percent of the data.

2 K-Nearest Neighbour

K-nearest neighbour (KNN) is a machine learning algorithm. This algoithms can be used for both classification and regression problems. In addition, (KNN) is supervised learning which means that we know the type of data our data is and what outcomes we are looking for. Futhermore, for implementation of the KNN algorithm we will refrence back to the PCA problem and use the 3 components based coordinates and other categorical variables which explain the 82 percent varience to do KNN.

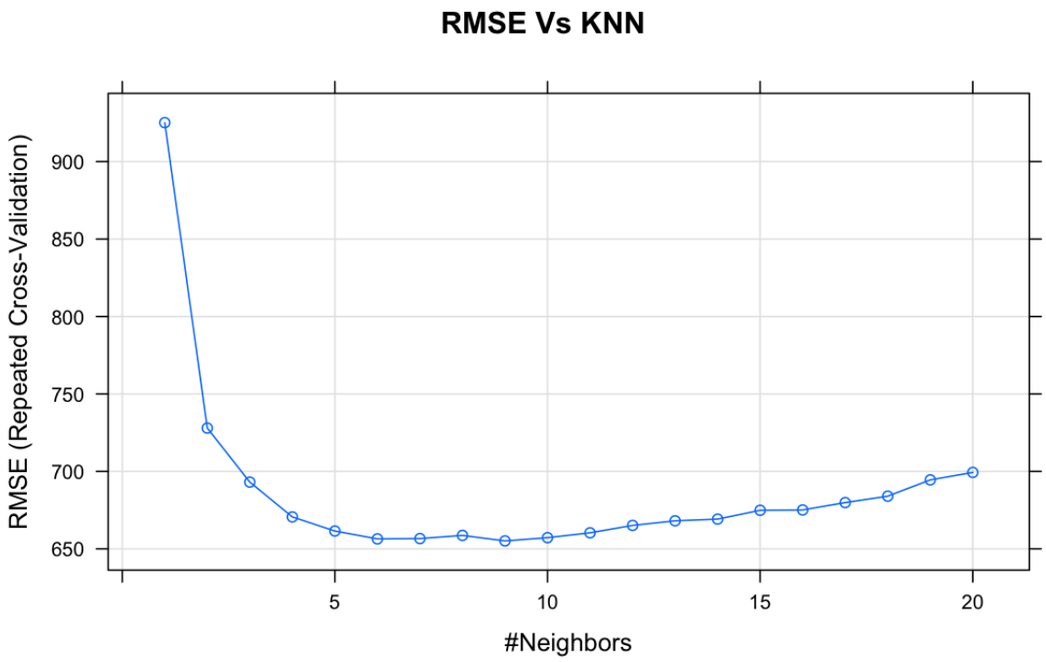

Figure 2: This figure show the average RMSE of the estimated prediction error on a validation set. The value of K for which the RMSE is lowest is the value of K that we will take. So that is K = 9.

k-Nearest Neighbors **512 samples** **10 predictor** **No pre-processing** **Resampling: Cross-Validated (10 fold, repeated 3 times)** **Summary of sample sizes: 460, 461, 460, 461, 461, 461, ...** **Resampling results across tuning parameters:**			
k	**RMSE**	**Rsquared**	**MAE**
1	925.1560	0.7839712	690.8140
2	727.9760	0.8608545	560.1111
3	693.0851	0.8758349	538.5740
4	670.6142	0.8858940	526.6748
5	661.4997	0.8910960	523.4297
6	656.4150	0.8942347	519.3416
7	656.6373	0.8953553	522.3512
8	658.6736	0.8959119	522.6202
9	655.1282	0.8980874	521.0988
10	657.1985	0.8981599	524.6721
11	660.3379	0.8978063	527.6320
12	665.1246	0.8967227	534.0605
13	668.0968	0.8965219	536.8265
14	669.2021	0.8964224	539.2323
15	674.8979	0.8959398	543.9940
16	675.1185	0.8968547	546.7142
17	679.8648	0.8962774	552.8564
18	684.0043	0.8959587	556.5736
19	694.5197	0.8933500	565.4170
20	699.3458	0.8931800	570.3835

Table 2: This table here shows a list of RMSE values. The RSME was used to select the optimal model using the smallest value. The final value used for the model was k = 9.

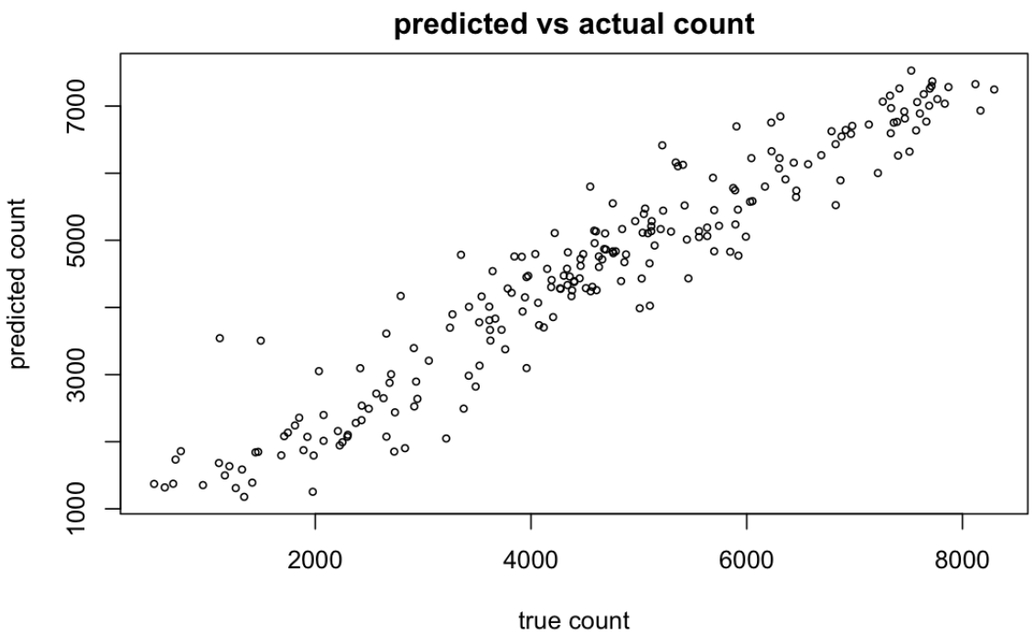

Figure 3: This figure show there's a strong correlation between the model's predictions and its actual results with our K = 9. With the The $R^2 = 0.912$ being very good since R^2 value close to 1 indicated a postive linear assosciation between the predicted vs actual data.