Team Project: Satellites

Statistical Learning

100364195, Alonso Beltrán 100459927, Gil Sénécaut
100469317, Eneko Lekuona

Contents

1 Introduction

1.1 Dataset

The chosen dataset for the team project includes entries for all publicly known satellites currently in orbit, compiled from different sources ranging from governmental agencies to sightings by particulars (*https://www.ucsusa.org/resources/satellite-database*).

1.2 Selected Variables

From the source database the following variables were selected:

- **Official name:** Official name of the satellite. Qualitative.
- **Purpose:** Purpose of the satellite (f.e.: Earth observation, technology development). Qualitative.
- **Orbit class:** General groupings of orbits (f.e.: low Earth orbit, LEO; geostationary orbit, GEO). Qualitative.
- **Perigee:** Distance from the surface of the Earth to the closest point of the orbit, in kilometers, illustrated in Figure 1.1. Continuous qualitative.
- **Apogee:** Distance from the surface of the Earth to the furthest point of the orbit, in kilometers, illustrated in Figure 1.1. Qualitative continuous.
- **Eccentricity** Eccentricity of the orbit. Ranging from 0 to 1, where 0 corresponds to a circular orbit, 1 to a parabolic trajectory and values greater than 1 to hyperbolic trajectories. The last two are not present in the database, since these trajectories would leave Earth's sphere of influence and could therefore not be considered to be Earth satellites. Related to apogee (r_{ap}) and perigee (r_{pe}) by

  $$e = \frac{r_{ap} - r_{pe}}{r_{ap} + r_{pe}}. \tag{1.1}$$

  Qualitative continuous.
- **Inclination** In degrees, inclination of the orbital plane with respect to Earth's Equatorial plane, illustrated in Figure 1.1. Qualitative continuous.
- **Period:** In minutes, the orbital period: the time taken by the satellite to complete an orbit. Related to apogee (r_{ap}) and perigee (r_{pe}) by

  $$T = 2\pi\sqrt{\frac{(r_{ap} + r_{pe})^3}{8GM}}, \tag{1.2}$$

  where G is the Universal Gravitational Constant and M is Earth's mass. Qualitative continuous.
- **Launch mass:** In kilograms, the mass of the satellite during launch.
- **Exp. lifetime:** Expected lifetime of the satellite, in years. Qualitative continuous.

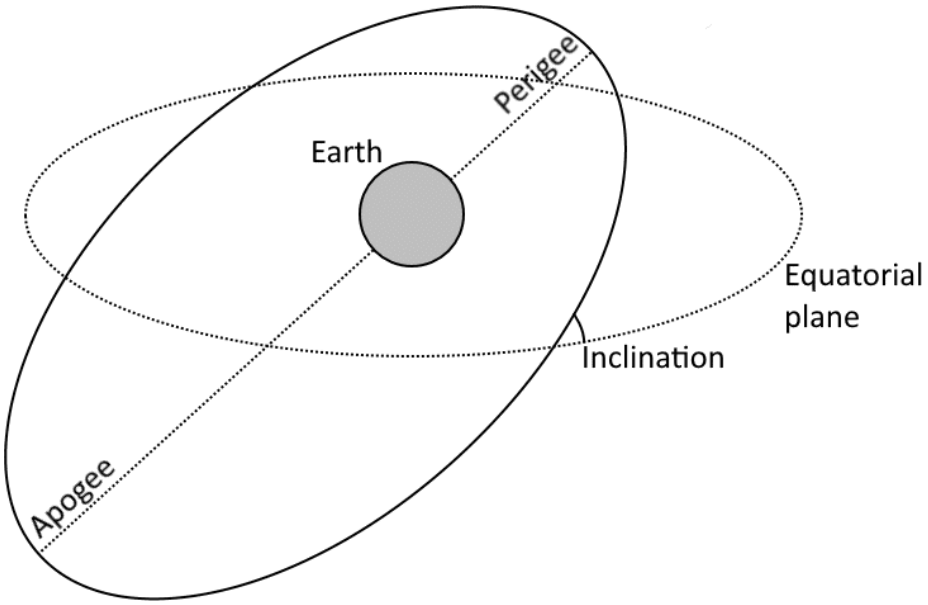
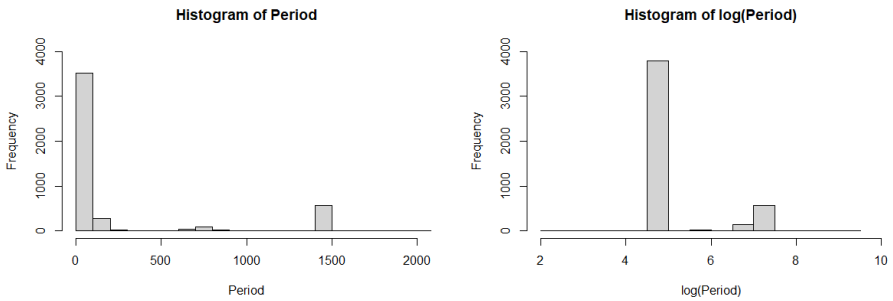


Figure 1.1: Schematic showing the definitions of perigee, apogee and inclination of an orbit around Earth.

2 Preliminary Analysis

2.1 Transformation of Quantitative Variables

Looking at the boxplots and histograms of each quantitative variable, we were able to determine that it would be helpful to obtain the logarithmic transformation of 5 out of the quantitative variables. In some instances, such as with the variable Period (Figures 2.1), the logarithmic transformation addresses the extreme skewness of the data.

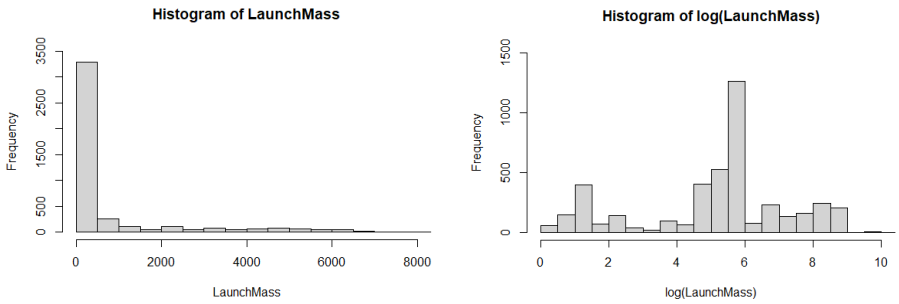


(a) Period without transformation.

(b) Period after logarithmic transformation.

Figure 2.1: Histograms of Period before (a) and after (b) logarithmic transformation.

In other cases, such as with the variable Launch Mass (Figures 2.2), the logarithmic transformation allows for our histogram to depict something that approximates a normal distribution.



(a) LaunchMass without transformation. (b) Launchmass after logarithmic transformation.

Figure 2.2: Histograms of LaunchMass before (a) and after (b) logarithmic transformation.

2.2 Imputation of Missing Data

Before any analysis of the dataset, we had already decided to omit some variables where only a large percentage of their data entries were missing. For instance, the variable DryMass had 4120 null entries (90.5%) and so it had to be omitted. Table 2.1 below is a table of how many null entries each of the selected variables have.

Table 2.1: Number of missing datapoints for each variable.

Variable	n_miss
`OfficialName`	0
`Purpose`	0
`OrbitClass`	0
`OrbitType`	0
`Eccentricity`	0
`Inclination`	0
`log_Perigee`	0
`log_Apogee`	0
`log_Period`	4
`log_LaunchMass`	236
`log_ExpLifetime`	1751

As we can see, most of the variables have zero null entries but the three quantitative variables Period, LaunchMass and ExpLifetime have 4, 236 and 1751 missing entries, respectively. R is the tool we used to assist us with the analysis of the dataset and with it we were able to impute all the missing data.

2.3 Analysis of Scatter Plot Matrix in terms of Purpose

Having made the necessary transformations to our data and imputations of the missing values, we looked at the scatterplot matrix of all the quantitative variables in the Satellites dataset in terms of our qualitative variable of interest, Purpose.

To reiterate, the entries of the variable Purpose classify all the satellites in the dataset in terms of their role. The most prevalent roles of these satellites are Communications (63%) and Earth Observation (22%).

Figure 2.3 below is a scatterplot matrix in terms of Purpose where the orange and green coordinates represent satellites whose role is Communications and Earth Observation, respectively. The blue coordinates represent the rest.

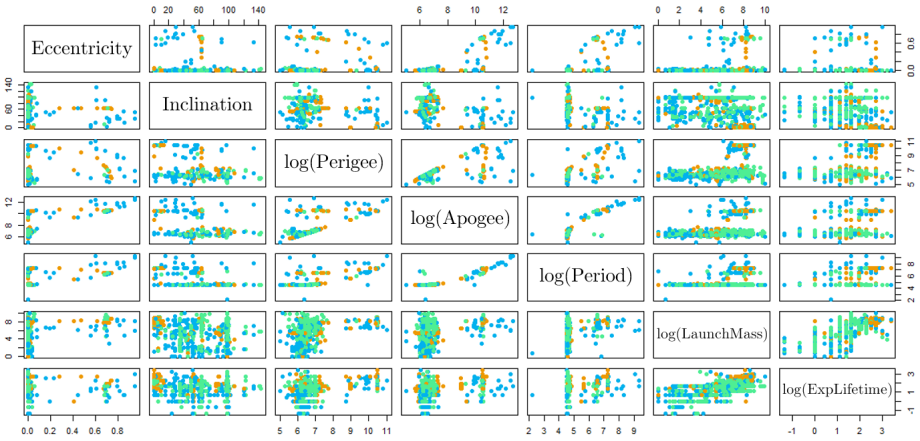


Figure 2.3: Scatterplot matrix colored by purpose. Green represents Earth observation satellites, orange communication satellites, and blue others.

This scatterplot matrix in Figure 2.3 allows us to extract a lot of information about our data, some expected, some surprising. Let us look at a few of these more closely.

2.3.1 Eccentricity and Inclination

We can extrapolate from our understanding of the terms eccentricity, apogee and period that satellites with very high eccentricity (close to 1) must have a very large apogee and period. As we can see from Figure 2.4 below, this certainly appears to be the case. What is also clear is that a very high concentration of satellites in this dataset have eccentricity equal to or very close to 0 which strongly suggests that the positive correlation between the variables must be very weak.

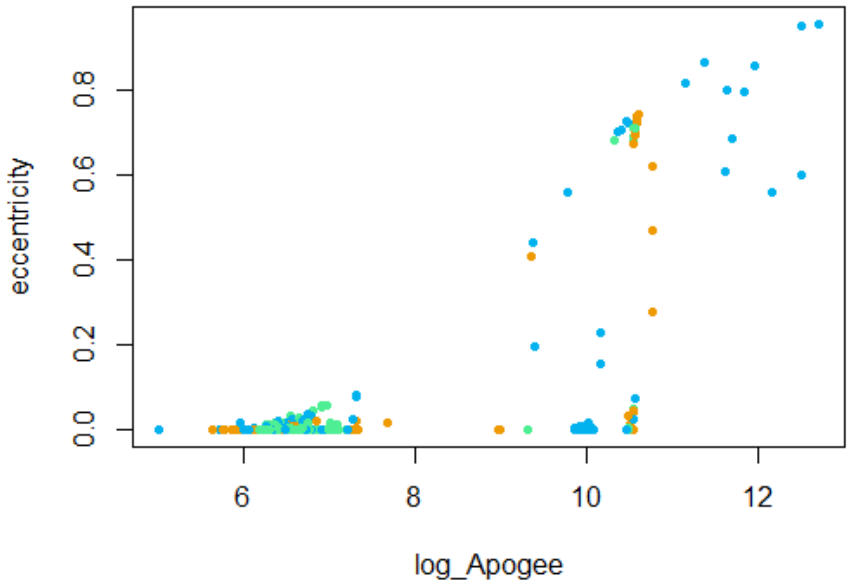


Figure 2.4: Scatterplot of eccentricity vs apogee, colored by purpose.

All the scatter plots that look at the variable inclination in relation to another variable do not reveal any correlation that inclination might have we any other quantitative variable. This leads us to assume that inclination is an independent variable. However, intuitively we suspect that a positive correlation between inclination and satellites in a low Earth orbit (low log_Perigee, log_Apogee, etc.) exists. This is because Earth Observation and Communications satellites in low Earth orbit require a larger inclination in order to cover more of the Earth's surface, while satellites with larger orbits do not have this requirement.

2.3.2 Perigee, Apogee and Period

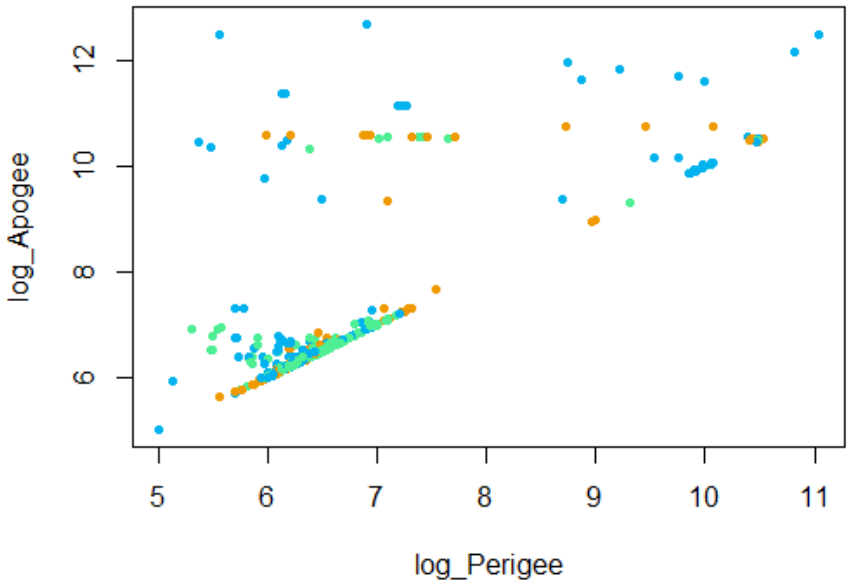
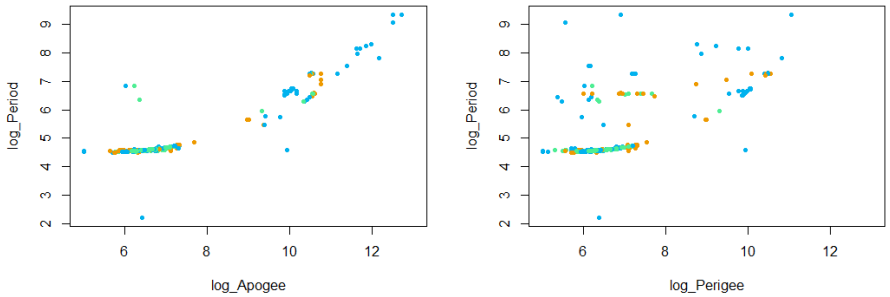


Figure 2.5: Scatterplot of apogee vs perigee, colored by purpose.

Since they represent the lowest and highest point of an orbit, by definition the perigee is always smaller than the apogee, and that is clearly represented in Figure 2.5. A majority of the satellites appear to have the same or very similar values for perigee and apogee, which is to say that they have a pretty circular orbits. This is reinforced by the concentration of satellites with eccentricity at or close to 0 that we can see in Figure 2.3. Considering orbital mechanics, it is not surprising that Figure 2.6 shows that both apogee and perigee have a strong positive correlation with the orbital period.



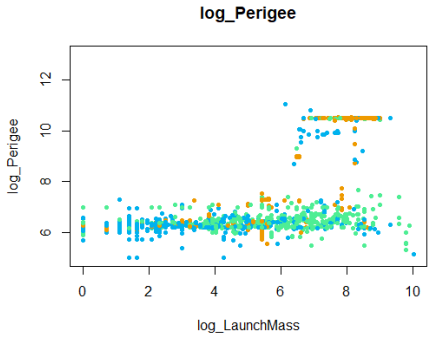
(a) Period vs apogee.

(b) Period vs perigee.

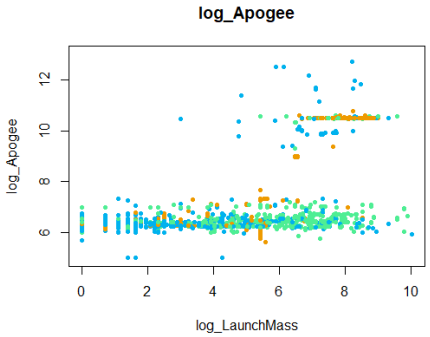
Figure 2.6: Scatterplots of period vs apogee (a) and perigee (b).

2.3.3 Earth Observation and Communications

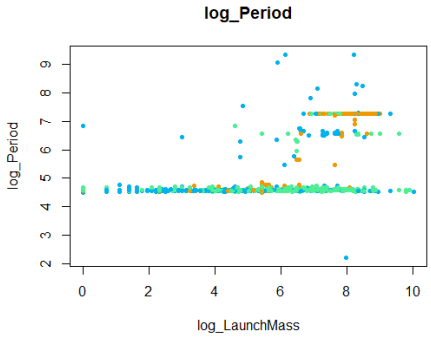
Although, it's unclear from Figures 2.4, 2.5 and 2.6, we can see from some of the scatter plots in Figure 2.3 that satellites with purpose Earth Observation and Communications form two distinct groups in the dataset. The clearest depiction of this can be seen in Figures 2.7.



(a) Scatterplot of perigee vs launch mass.



(b) Scatterplot of apogee vs launch mass.



(c) Scatterplot of period vs launch mass.

Figure 2.7: Scatterplots of perigee (a), apogee (b) and period (c) vs launch mass.

When it comes to the purpose of the satellites, most of those that are dedicated to Earth observation (green) are in low orbit (small perigee, apogee and period) in order to obtain higher resolution data per surface area. From Figure 2.7 we can also see that Earth observation satellites can vary a lot in mass. There are very low mass Earth observation satellites like the cubesat BisonSat with log_LaunchMass = 0, but high mass Earth observation satellites such as the Resurs-P1 (log_LaunchMass = 8.68) can also be found in a low Earth orbit. Most communication satellites (orange), however, look for higher orbits (large perigee, apogee and period) to cover more area with a single satellite. While satellites of all masses exist in lower orbits, only massive satellites can get to the high orbits that communication satellites reach.

2.3.4 Launch Mass and Expected Lifetime

Larger mass also appears to strong correlation with a longer expected lifetime and since we established that communication satellites are for the most part high mass satellites, we can see in Figure 2.8 that they are all bunched up in the top right of the scatter plot of these two variables.

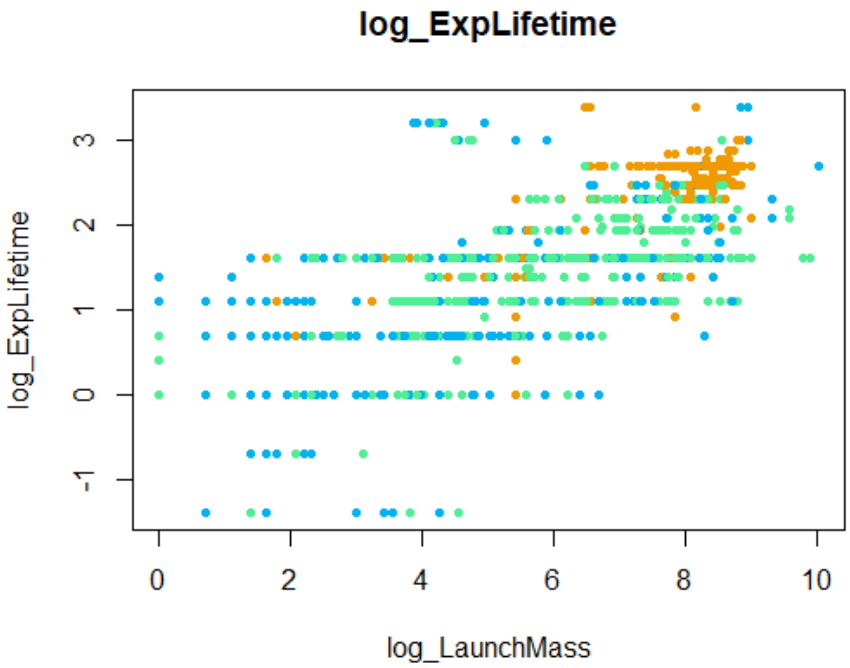


Figure 2.8: Scatterplot of expected lifetime vs launch mass.

2.4 Estimating the Mean Characteristics

Using R, we were able to obtain the sample mean vector (Table 2.2), covariance matrix (Table 2.3) and correlation matrix.

Table 2.2: Sample mean vector.

Variable	m_quan
`Eccentricity`	0.009006144
`Inclination`	61.882918681
`log_Perigee`	6.961140218
`log_Apogee`	7.029421237
`log_Period`	4.984873801
`log_LaunchMass`	5.108706750
`log_ExpLifetime`	1.489632189

Table 2.3: Covariance matrix.

S_quan	Ecc.	Inc.	Pe.	Ap.	Period	L.M.	Exp.LT.
Ecc.	0.005	-0.096	0.003	0.030	0.016	0.014	-0.000
Inc.	-0.096	969.624	-30.608	-30.838	-20.696	-41.914	-11.543
Pe.	0.003	-30.608	2.233	2.213	1.373	1.909	0.681
Ap.	0.030	-30.838	2.213	2.337	1.436	1.949	0.670
Period	0.016	-20.696	1.373	1.436	0.923	1.215	0.418
L.M.	0.014	-41.914	1.909	1.949	1.215	4.956	1.204
Exp.LT.	-0.000	-11.543	0.681	0.670	0.418	1.204	0.501

Figure 2.9 is a helpful visualization of the correlation matrix where the size and shade of the circles indicate the strength of the correlation and the color distinguishes between a positive and negative correlation. For instance, a large dark blue circle reveals a strong positive correlation.

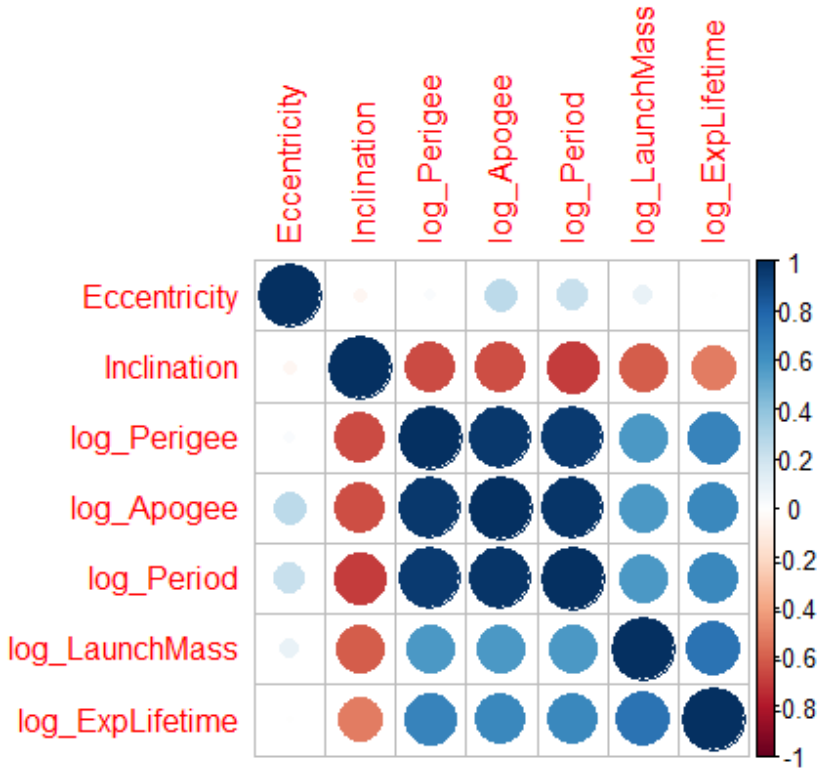


Figure 2.9: Correlation plot.

The most interesting information can be found in Figure 2.9 as the information it gives about our dataset is in accordance with predictions that we made about these quantitative variables based on graphical analysis in Section 2.3. The first evidence of this is regarding eccentricity along with apogee and period. Figure

2.4 revealed that a vast majority of the satellites in this dataset have eccentricity equal to or very close to 0 (96.5% have eccentricity less than 0.01). This allowed us to determine that the positive correlation between eccentricity and each of apogee and period would be very weak and that all other variables would have essentially no correlation with eccentricity, both of which can clearly be seen in Figure 2.9. As we suspected, there also exists a negative correlation between inclination and high Earth orbit (high log_Perigee, ..., high log_ExpLifetime).

Furthermore, it was expected by definition and from the scatter plots we looked at in Section 2.3.2 that the variables perigee, apogee and period are heavily positively correlated, which we can clearly see is true from Figure 2.9. By inspection of the scatter plot matrix and individual scatter plots in Sections 2.3.3 and 2.3.4, we were also able to correctly predict a positive correlation between the variables period and launch mass as well as launch mass and expected lifetime. By deduction we know that perigee and apogee are positively correlated to launch mass and expected lifetime as well, and this is reflected in Figure 2.9.

2.5 Outliers

Ideally, we would want to rid the dataset of any potential outliers in our dataset to avoid any distorted analysis of the data. Thus, we used R to compute the Minimum Covariance Determinant (MCD) estimators. It quickly became clear that this method would not work with our dataset.

The method of computing the MCD estimators is not appropriate for asymmetric datasets and although we transformed most of the variables to address their extreme skewness, our dataset can very clearly be split up into two groups as we saw in section 2.3.3 which unfortunately leads to significant interference.

As can be seen in Figure 2.10 below, removing the outliers from our dataset with the MCD method was rather uninformative as it only led to a very slight strengthening of the correlations that all variables have with each other, except regarding eccentricity. In fact, with eccentricity we see a change in the opposite direction of what we were expecting. It is very likely that due to the distribution of satellites in terms of eccentricity, the satellites that had high eccentricity were wrongly classified as outliers and this somehow led to eccentricity having a weak negative correlation with all other variables except inclination and a weak positive correlation with inclination itself.

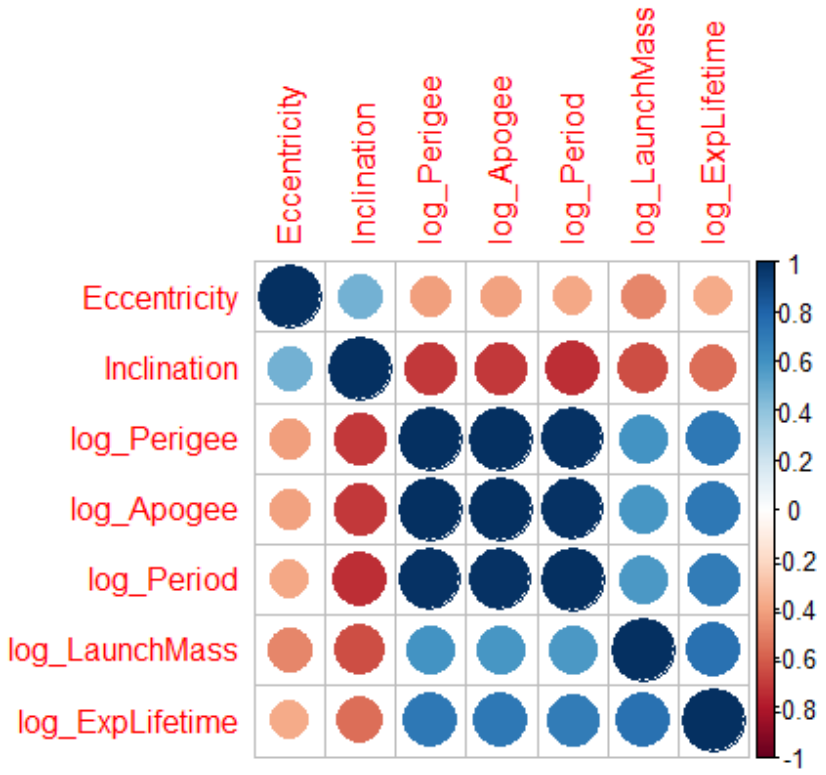


Figure 2.10: Correlation plot after removing the outliers with MCD.

3 Dimension Reduction Techniques

The main goal of this chapter will be the implementation of dimension reduction techniques, which will reveal the existence of groups and outliers as well as reducing the number of variables while preserving most of the information.

3.1 Principal Component Analysis

Principal Component Analysis (PCA) only works with quantitative variables, and it is as well a method that should be used with standardization of the variables as the variables have different units of measurement. In order to achieve this equal effect of the variables, we set the scale argument to TRUE, which sets the standard deviations of all variables to be equal to 1.

How many principal components (PCs) should be selected? Even though PCA returns as many PC's as variables (7 in this case), the point of PCA is to reduce dimensionality, so we will concentrate on the most relevant principal components. To do this, we calculate the percent of total variance explained by each principal component, shown in Table 3.1.

Table 3.1: Importance of components.

	Std. deviation	Prop. of variance	Cumulative Prop.
PC1	2.1247	0.6449	0.6449
PC2	1.0350	0.1530	0.7979
PC3	0.8525	0.1038	0.9017
PC4	0.67852	0.06577	0.96750
PC5	0.44633	0.02846	0.99595
PC6	0.15968	0.00364	0.99960
PC7	0.0531	0.0004	1.0000

In order to show the proportion of variance for each principal component we make use of the scree plot in Figure 3.1.

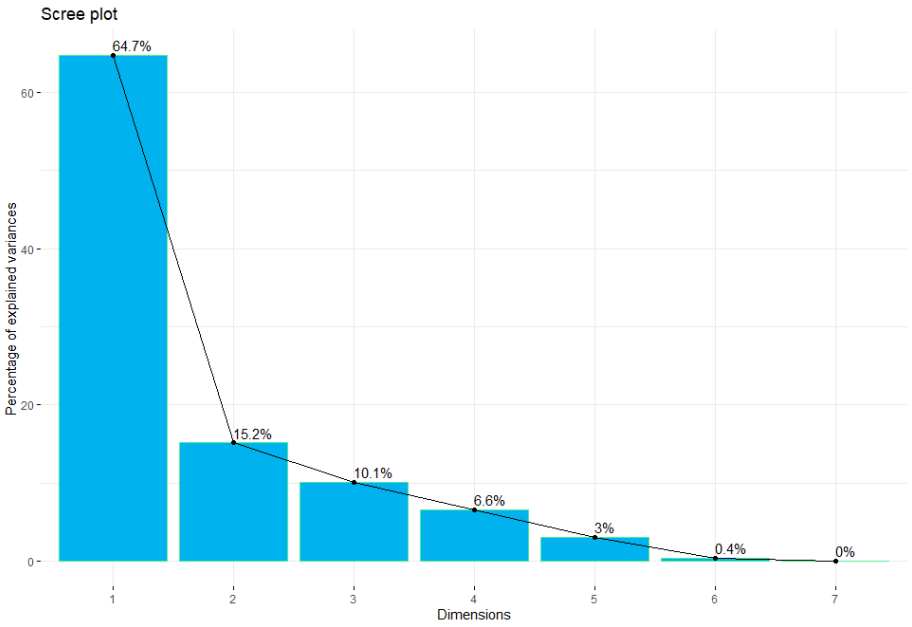


Figure 3.1: Proportion of variance for each principal component.

As we can see in Table 3.1 and in the scree plot (Figure 3.1), two principal components are enough, due to the fact that they contain 80% of the total variance. In this case, we reduce the dimension of the data set from 7 to 2. That is, we keep 28.5% of the number of variables in the original data which will account for around the 80% of the variance inside.

If we pay attention in the interpretation of these principal components and their weights, we can understand the importance of each variable.

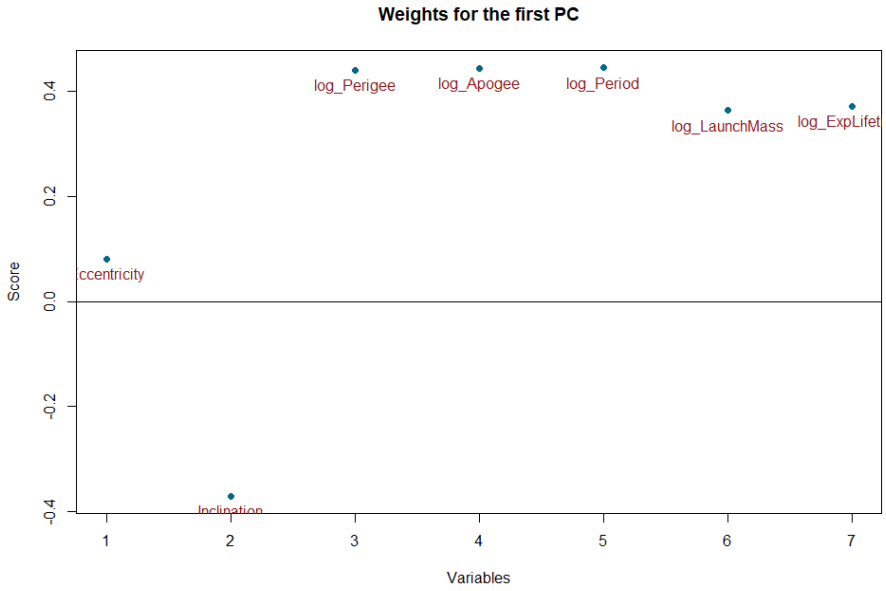


Figure 3.2: Weights for first principal component.

As illustrated in Figure 3.2, PC1 correlates similarly to several variables. We see that it has positive weighting for perigee, apogee, period, launch mass and expected lifetime, and a negative weighting for inclination. The interpretation of this variable taking into account the loadings is that this PC is a measure of large satellites in high altitude low inclination orbits.

PC2 correlates most strongly with Eccentricity, as shown in Figure 3.3. In fact, we could state that based on the correlation of 0.916, this principal component is primarily a measure of the Eccentricity as seen in the next figure:

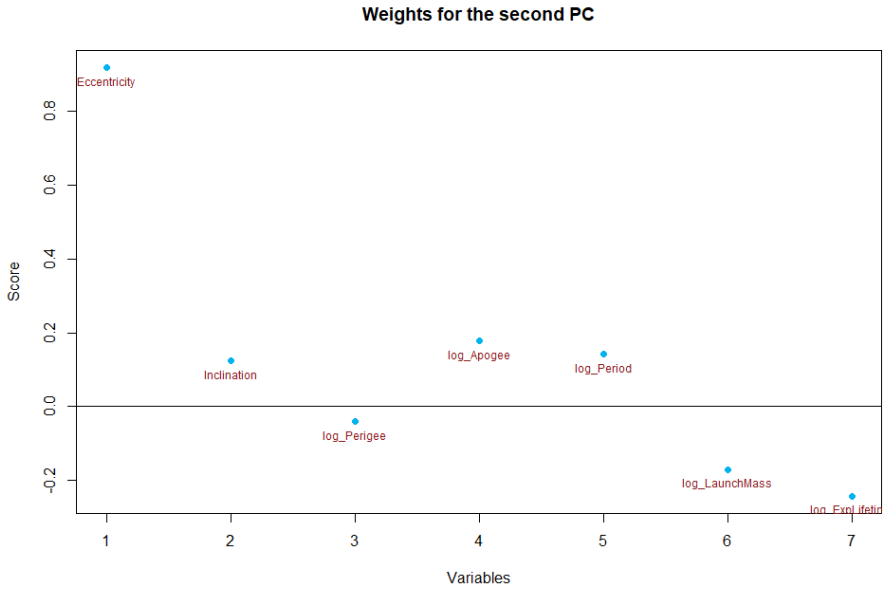


Figure 3.3: Weights for second principal component.

In order to have a representation of the loadings and scores in a single plot we can make use of the biplot (Figure 3.4). In one hand, we have the vectors that represent the variables. In this case, they hold a similar length, telling us that each of their information are similarly expressed. Furthermore, taking a look to the angles, we can reaffirm what we saw in the previous figures, that all variables but eccentricity are close to be parallel to PC1, meaning that they are contributing to its creation, bearing between them small angles (high correlation between them) and almost only eccentricity produces PC2, which has an angle close to 90 degrees to the other variables (it is uncorrelated to them). In the other hand, the scores are represented showing us that indeed there are 3 differentiated groups.

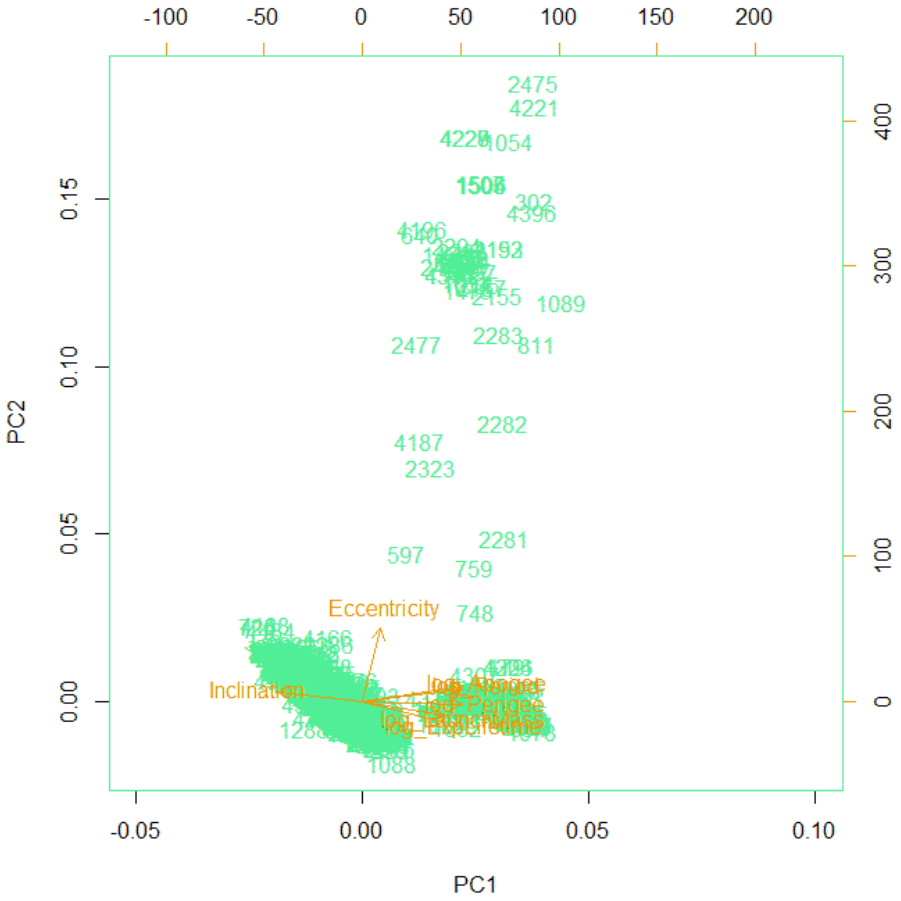


Figure 3.4: Biplot representing scores and loadings.

The correlation plot in Figure 3.5 summarizes nicely the importance of each original variable for each principal component, making clear the importance of the first two.

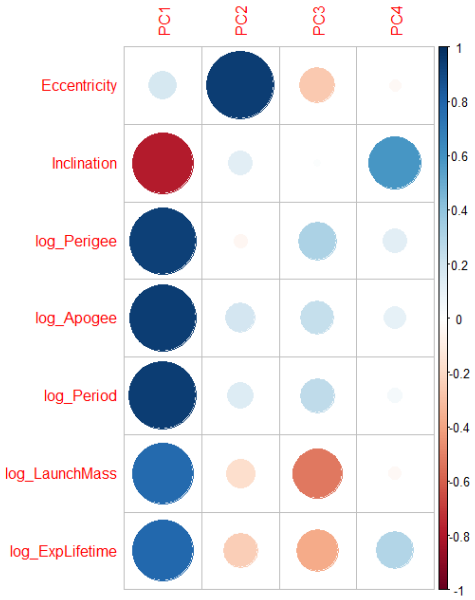


Figure 3.5: Correlation plot of first four principal components.

In the scatter plot matrix in Figure 3.6 some of the pairs show the presence of outliers and two clear groups, but only PC1 vs PC2 highlights the existence of three different groups. The green, orange and blue groups are satellites with the purpose of "Earth Observation", "Communications", and "others" respectively.

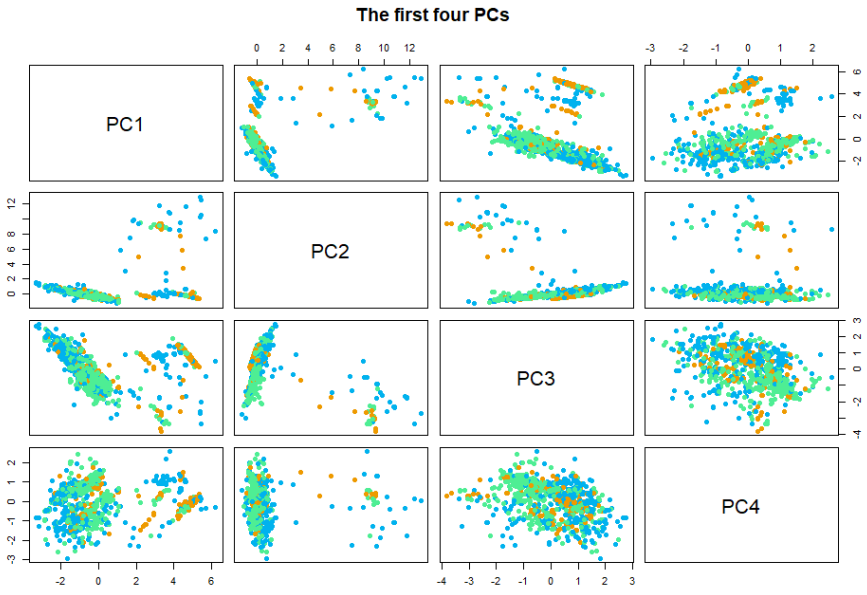


Figure 3.6: Scatter plot matrix of the first four principal components.

3.2 Independent Component Analysis

Independent component analysis is an alternative and perhaps a more complex technique, based on the idea of maximizing the statistical independence and non gaussianity of the variables.

After obtaining the independent components (ICs) and ordering them by measuring their non gaussianity with neg-entropy, we obtain Figure 3.7, which shows us that there are two ICs with much larger negative entropy than the others, meaning that they are highly non Gaussian.

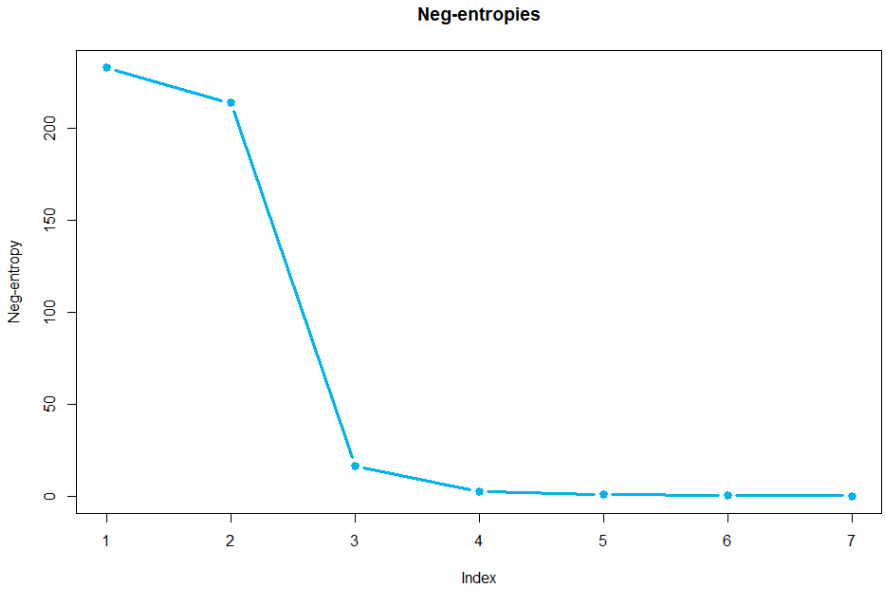


Figure 3.7: Representation of negative entropies for each independent component.

Obtaining the scatter plot matrix with the same purpose colors we set in PCA, the results are those shown in Figure 3.8.

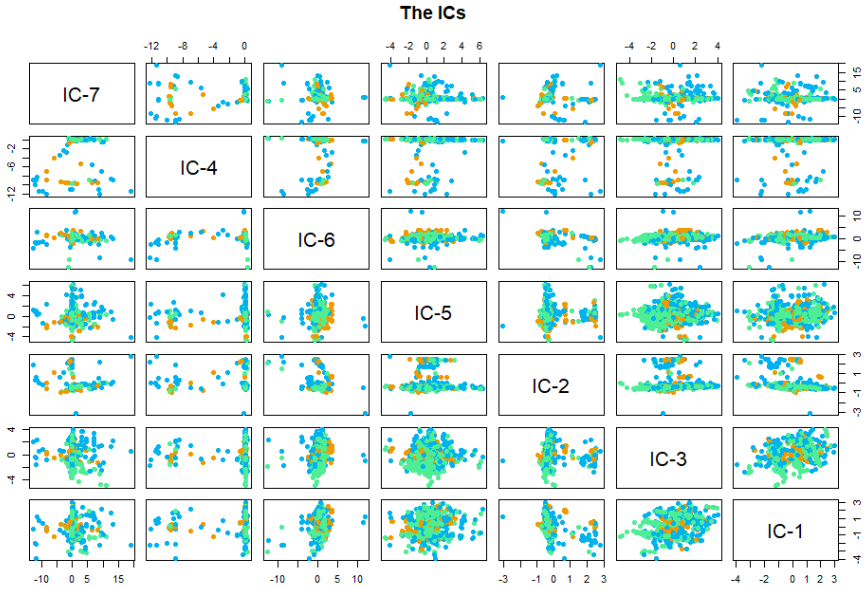


Figure 3.8: Scatterplot matrix of independent components.

In some cases, the presence of groups is only revealed in pairs of ICs with low negative entropy, due to the fact that data skewness is almost equal to zero (very symmetrical), and therefore holding a very small kurtosis coefficient. However, this is not a strict and universal rule. As we can see from Figure 3.8, IC-2 shows some distinction between groups and has low neg-entropy, nevertheless, something similar occurs in the column of IC-4, which has larger negative entropy.

The IC with the greatest number of variables related is IC-2, as shown in Figure 3.9, being one of the ICs that make a better job separating groups (along with IC-4). It is normal that ICs with smaller negative entropies are related with several variables when the dataset has high asymmetry or holds extreme values. In this case IC-2 is one of the ICs with lowest negative entropy, alongside IC-3 and IC-1, as seen in Figure 3.7.

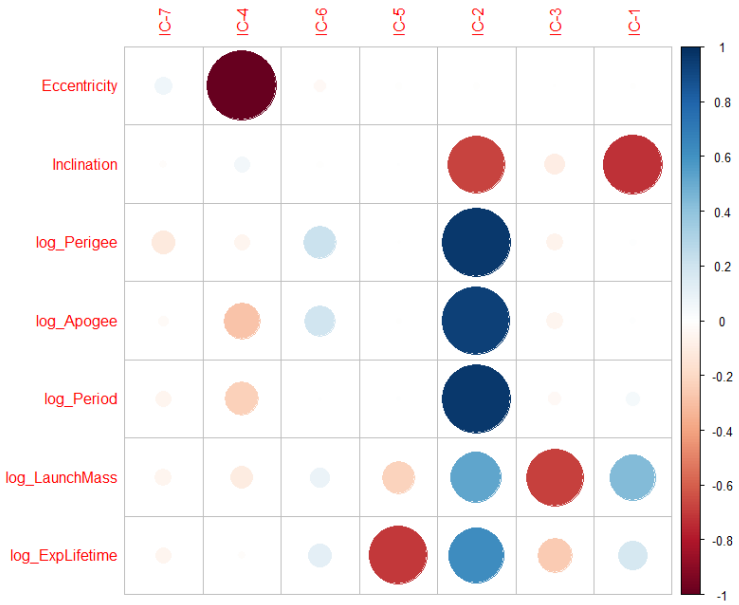


Figure 3.9: Correlation plot of independent components.

Plotting the PCs vs ICs in Figure 3.10, it can be appreciated how some of the ICs are highly correlated with the PCs, so at some point they give similar information, although in principle it is not a necessary thing to happen.

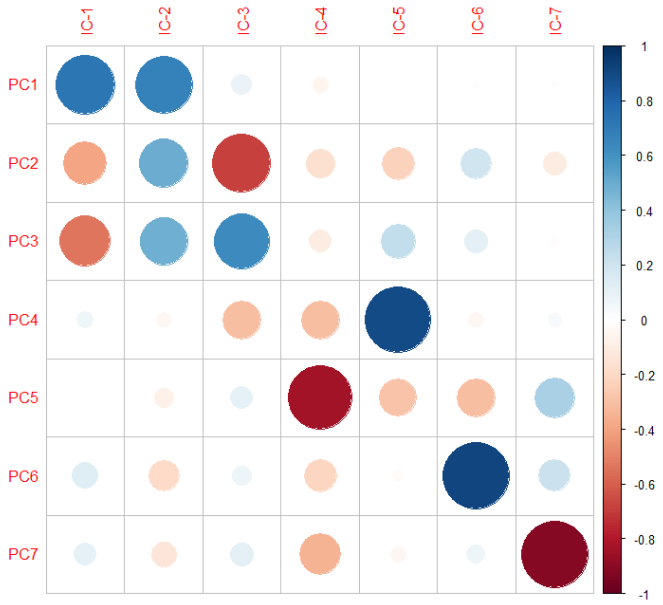


Figure 3.10: Correlation plot of independent components vs principal components

Both PCA and ICA are useful to highlight the presence of outliers and groups, although for this particular dataset it seems that PCs separate the groups in a clearer way than ICs, something that is completely normal and that is why it is useful to apply both methods and check which one is the one that suits the dataset better.

4 Unsupervised Classification

The goal of unsupervised classification is to divide the dataset in homogeneous groups or clusters. This can be achieved by a number of methods following different criteria, which can yield differing results. In this section we explore three of these methods to cluster our dataset using only the quantitative variables.

4.1 Partitional Clustering

Partitional clustering requires some original clusters to begin with, and works by exchanging points between these until a stable configuration is found. A popular partitional clustering method is K-means clustering, where in each step of the execution each point is assigned to the closest mean point of the clusters in the previous steps.

4.1.1 Optimal k Value

K-means requires the number of clusters k as an input parameter, which means that an optimum value must be found. A way to achieve this is to plot the total within sum of squares after the clustering has been performed with a range of values of k, called elbow graph because of its characteristic shape. In Figure 4.1, the elbow can be visually located in $k = 3$, where the slope of the curve sharply drops and adding more clusters has diminishing returns.

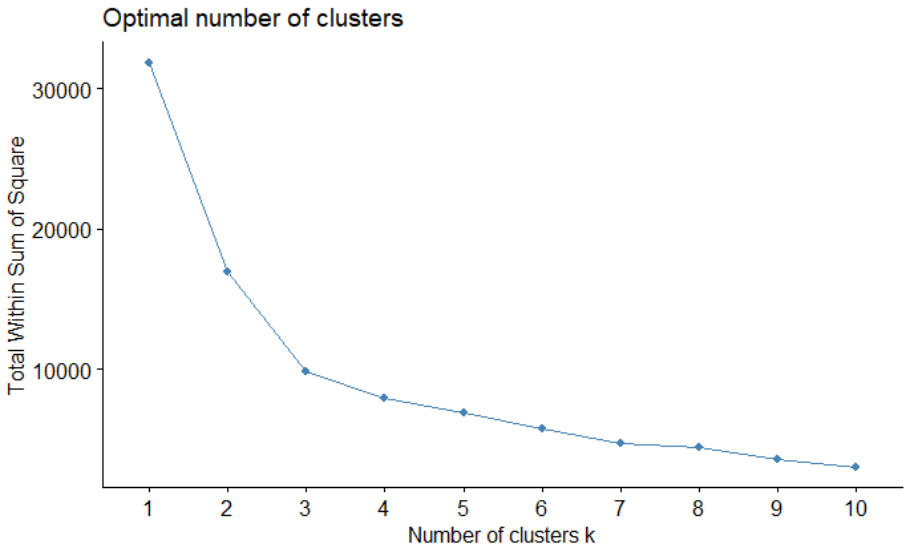


Figure 4.1: Elbow graph constructed using K-means with k ranging from 1 to 10. The elbow can be located at $k = 3$.

The silhouette is another way to determine the optimal value of k. For each point, the silhouette gives a measure of how well it is matched to its own cluster. Ranging from -1 to 1, negative values indicate a poor match while positive values show a good match. Figure 4.2 shows the average silhouette of the clustering with K-means with k ranging from 1 to 10. The highest average silhouette and therefore the optimum value is at $k = 3$, in agreement with the elbow plot.

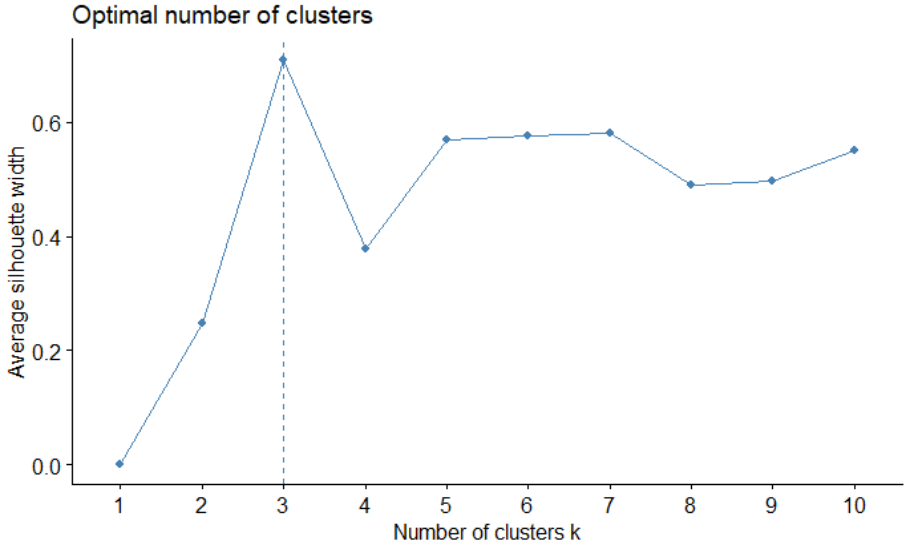
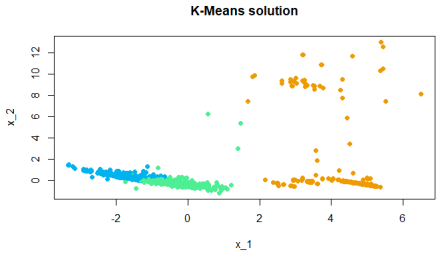


Figure 4.2: Plot of the average silhouette width constructed using K-means with k ranging from 1 to 10. The vertical line indicates the optimal value at $k = 3$.

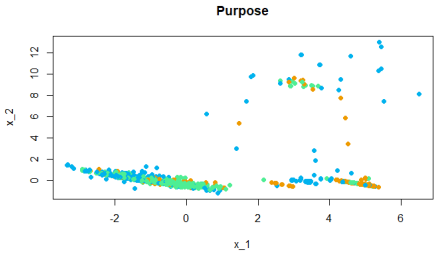
In a similar way, the gap statistic can also be used to find the optimum value of k. This is a measure of the difference between the distance within clusters to its expected value if there were no groups, where the optimal is found when the gap statistic is maximized. Unfortunately, the R functions to calculate the gap statistic failed to converge with our dataset, so we were forced to exclude this value from our analysis.

4.1.2 Comparison to Categorical Variables

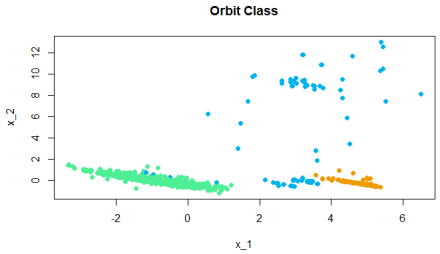
With the elbow graph and the silhouette in agreement with its optimal value for our dataset, performed the grouping using K-means with $k = 3$. The results of this grouping in the first two principal components are shown in Figures 4.3.



(a) Groupings by K-means.



(b) Groupings by purpose. Green dots indicate Earth Observation satellites, orange dots communication satellites and blue satellites with other purposes.



(c) Groupings by Orbit class. Green dots indicate satellites in low Earth orbit, orange dots satellites in geostationary orbits and blue satellites in other orbits.

Figure 4.3: Scatter plots of the three principal components grouped by K-means (a), purpose (b), and orbit class (c).

As Figures 4.3 show, at least by looking at the first two principal components in our dataset, the clustering performed with K-means (Figure 4.3a) follows roughly the distribution by satellite purpose (Figure 4.3b), our qualitative variable of interest. However, this qualitative variable being hard to predict just with orbital data, the predicted clusters have much clearer boundaries than reality.

4.2 Hierarchical Clustering

Unlike in partitional clustering, hierarchical clustering methods do not require a set number of clusters into which to divide the datapoints. Instead, they either start with as many clusters as datapoints with one in each and progressively combine them, or start with a single cluster with all the datapoints and divide them. To determine when to combine or divide clusters, they require a measure of the distance between clusters.

The different methods to determine the distances between clusters are called linkage methods. These are single linkage, complete linkage, average linkage and ward linkage. Since none is better than the others for all cases, all four linkage methods were studied to determine the most appropriate for our case, using the Manhattan distance.

4.2.1 Single Linkage

Single linkage considers the distance between clusters to be the minimum distance between points from one cluster to points in the other. The dendogram in Figure 4.4 shows how this method divides our dataset when we cut off to the point with $k = 3$ clusters, using the value obtained in partitional clustering.

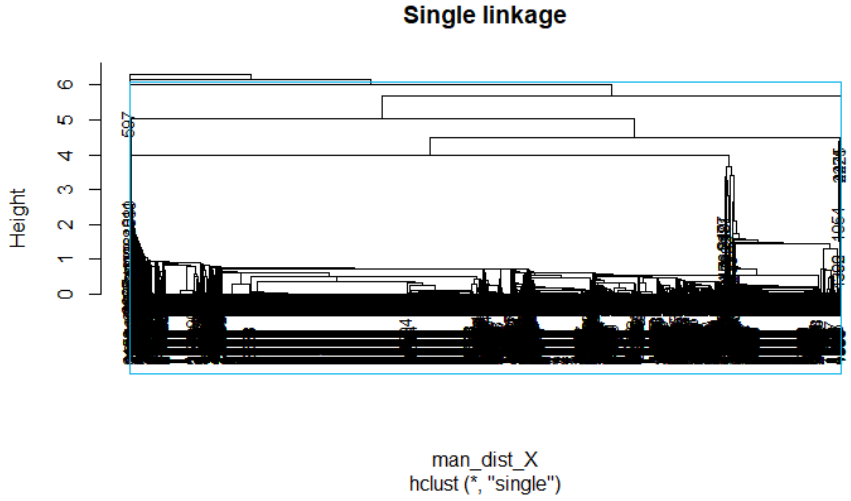


Figure 4.4: Dendogram of single linkage. Boxes show partition with $k = 3$.

Single linkage has grouped virtually all datapoints (4547) into the same cluster with only 1 in the second and 2 in the third. This is obviously not an acceptable clustering, which can be clearly seen in the silhouette plot in Figure 4.5.

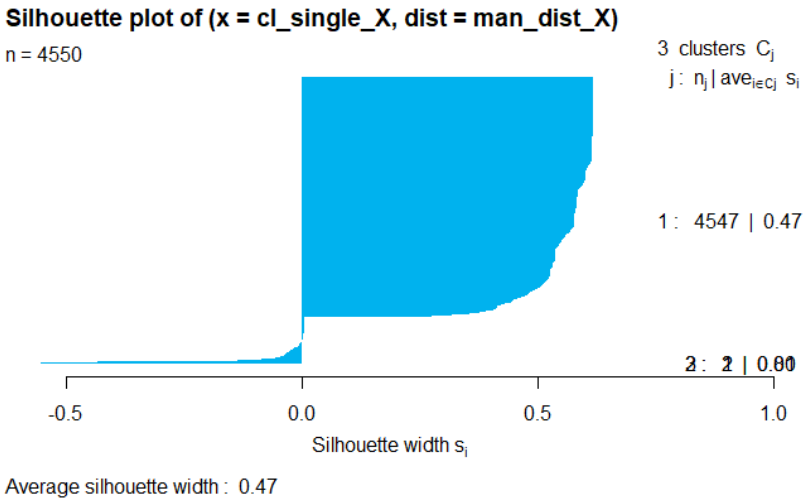


Figure 4.5: Silhouette plot of single linkage.

4.2.2 Complete Linkage

Complete linkage considers the distance between clusters to be the maximum distance between points from one cluster to points in the other. The dendogram in Figure 4.6 shows how this method divides our dataset when we cut off to the point with $k = 3$ clusters.

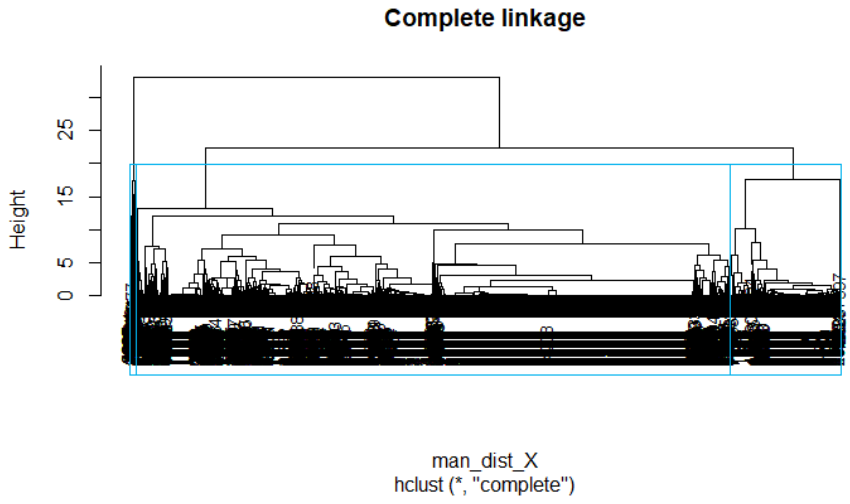


Figure 4.6: Dendogram of complete linkage. Boxes show partition with $k = 3$.

Complete linkage has divided the clusters more uniformly than single linkage, but a cluster still contains the mayority of the datapoints (3799), while another has very few (45) and the last groups a sizeable minority (706). The silhouette

plot in Figure 4.7 shows improvement from single linkage, but with some clearly mismatched points in cluster 3.

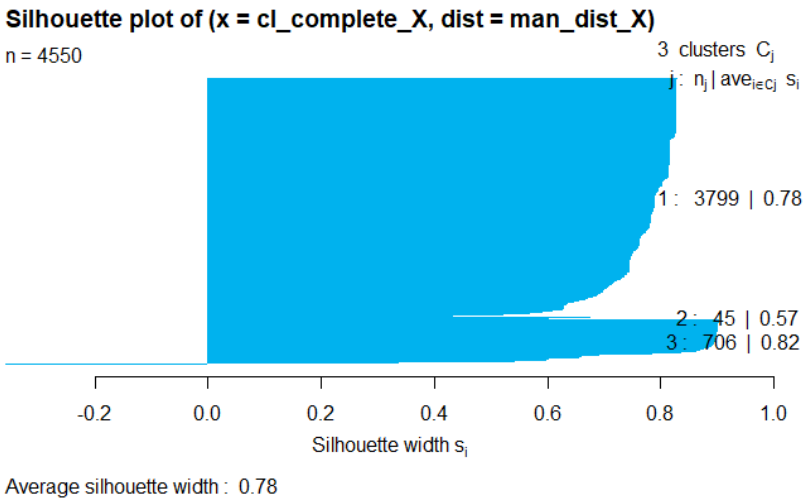


Figure 4.7: Silhouette plot of complete linkage.

4.2.3 Average Linkage

Average linkage considers the distance between clusters to be the average distance between points from one cluster to points in the other. The dendogram in Figure 4.8 shows how this method divides our dataset when we cut off to the point with $k = 3$ clusters.

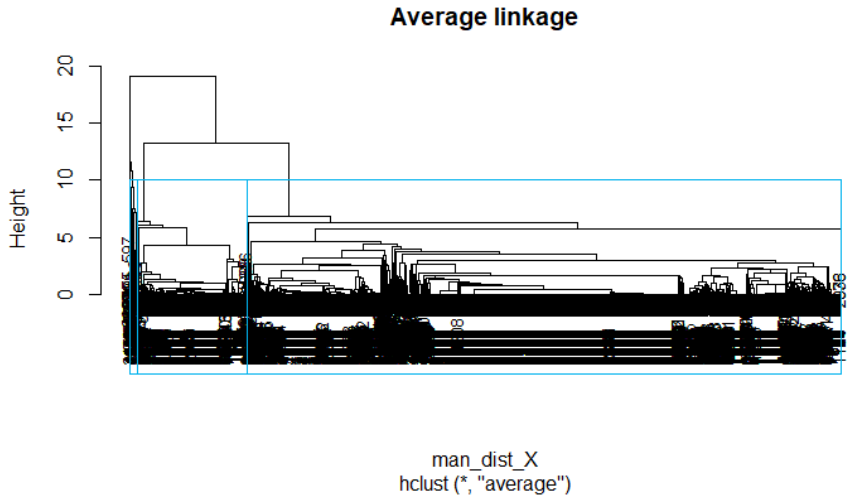


Figure 4.8: Dendogram of average linkage. Boxes show partition with $k = 3$.

Average linkage groups our dataset in a very similar manner to complete link-

age, at least when it comes to cluster size, with clusters with 3799, 48, and 703 datapoints. The silhouette, shown in Figure 4.9, shows improvement upon complete linkage with no clearly mismatched points.

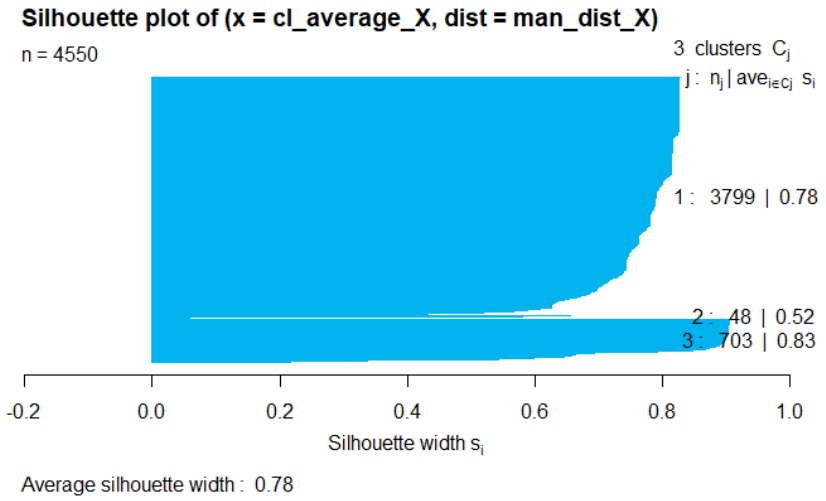


Figure 4.9: Silhouette plot of average linkage.

4.2.4 Ward Linkage

Ward linkage considers the distance between clusters to be the squared Euclidean distance between their sample mean vectors. The dendogram in Figure 4.10 shows how this method divides our dataset when we cut off to the point with $k = 3$ clusters.

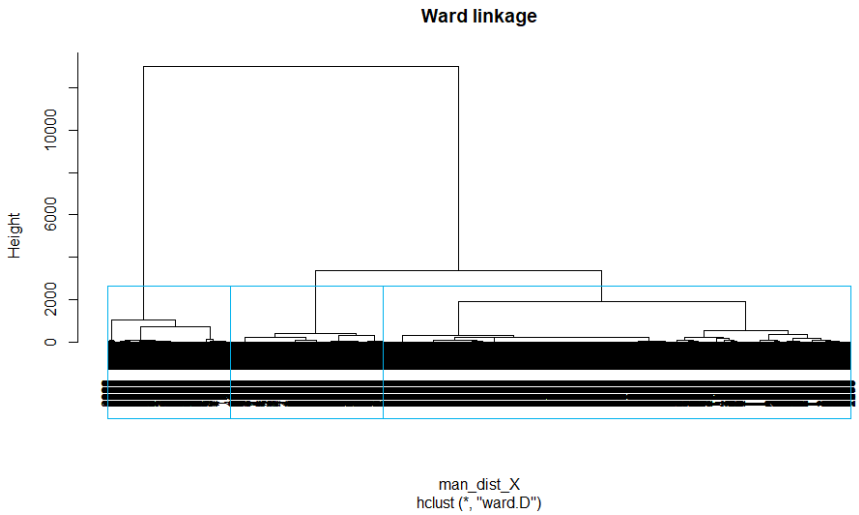


Figure 4.10: Dendogram of ward linkage. Boxes show partition with $k = 3$.

Although ward linkage groups our dataset more uniformly (groups of size 936, 751, and 2863), the silhouette in Figure 4.11 indicates worse grouping than both complete and average linkage.

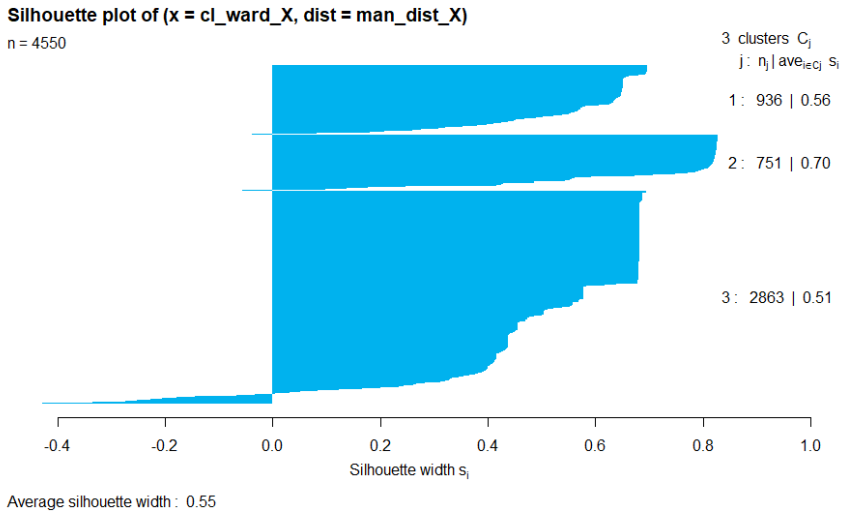
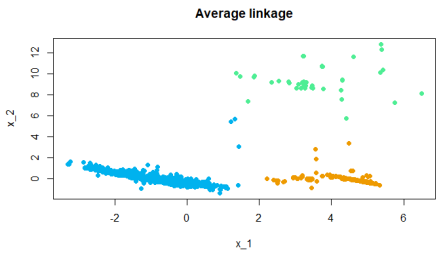


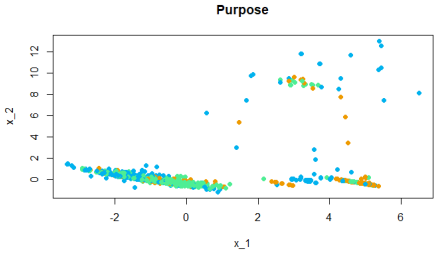
Figure 4.11: Silhouette plot of ward linkage.

4.2.5 Comparison to Categorical Variables

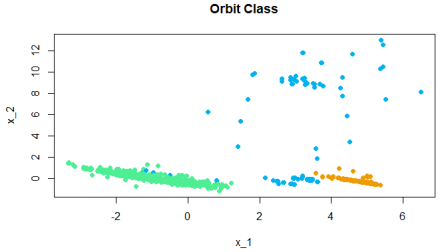
Since average linkage was the hierarchical clustering method that yielded the best results for our dataset, its grouping was compared to the groupings by purpose (the categorical variable of interest) and orbit class. As Figures 4.12 show, unlike partitional clustering which was closer to purpose, hierarchical clustering with average linkage appears to have grouped the satellites in a similar manner to orbit class.



(a) Groupings by average linkage.



(b) Groupings by purpose. Green dots indicate Earth Observation satellites, orange dots communication satellites and blue satellites with other purposes.



(c) Groupings by Orbit class. Green dots indicate satellites in low Earth orbit, orange dots satellites in geostationary orbits and blue satellites in other orbits.

Figure 4.12: Scatter plots of the three principal components grouped by average linkage (a), purpose (b), and orbit class (c).

4.3 Model Based Clustering

Unlike partitional and hierarchical clustering methods, model based clustering does not group by distance between datapoints. Instead, it creates the cluster with probabilistic methods assuming that the different clusters originate from different distributions with their own probabilities.

To determine the optimal number of clusters and their distributions, the Bayesian Information Criterion (BIC) is used, the model with the lowest BIC being the best one. Figure 4.13 and the table below show the BIC for different models with the number of clusters ranging from 1 to 10. Unlike of what we saw in partitional clustering, this criterion considers 7 to be the optimal number of clusters.

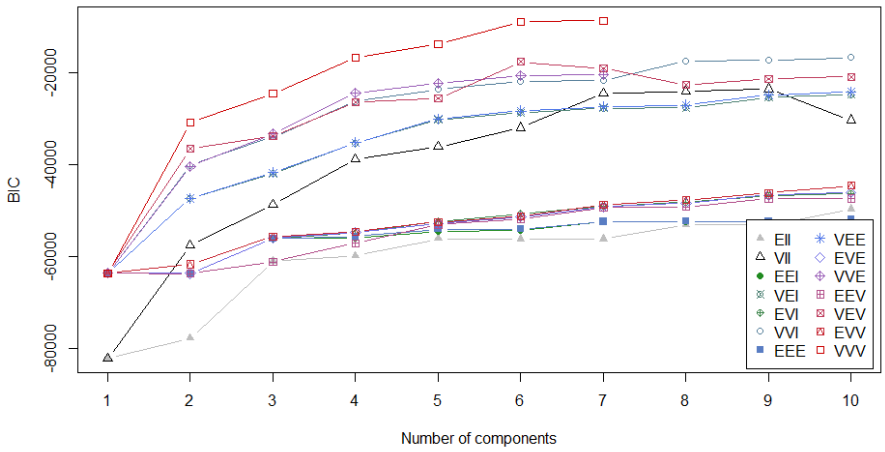


Figure 4.13: BIC, with the number of clusters ranging form 1 to 10.

```
> BIC_X
Bayesian Information Criterion (BIC):
         EII       VII       EEI       VEI       EVI
1  -82199.04 -82199.04 -63599.68 -63599.68 -63599.68
2  -77984.03 -54729.84 -63628.47 -47288.97 -63615.79
3  -61000.42 -46040.82 -56039.33 -41812.12 -56020.73
4  -59790.17 -35991.31 -55941.16 -37804.39 -54786.38
5  -56064.31 -34346.57 -56635.62 -33799.37 -52338.92
6  -56080.59 -28623.42 -53141.32 -33054.24 -50594.55
7  -50455.12 -28318.98 -53166.59 -21979.45 -49489.93
8  -56110.49 -23272.20 -52365.43 -19308.72 -48123.14
9  -50505.66 -23677.39 -52342.29 -19135.09 -46959.98
10 -50530.92 -23214.87 -52367.58 -17006.37 -46989.34

          VVI       EEE       VEE       EVE       VVE
1  -63599.682 -63608.10 -63608.10 -63608.10 -63608.10
2  -36370.256 -63637.66 -47286.41 -63618.95 -36522.48
3  -30250.776 -55984.75 -41672.09 -55955.21 -29349.55
4  -22677.403 -57874.02 -37806.53 -54711.42 -20602.74
5  -20198.751 -56499.87 -33803.92 -52025.91 -23457.34
6  -15111.247 -52872.29 -33057.33 -50582.50 -13940.11
7  -15737.672 -52897.64 -21782.51 -49494.83 -14627.13
8  -11305.349 -52364.20 -18922.75 -48100.02        NA
9   -9861.346 -52334.52 -18462.15 -46286.70        NA
10  -7790.461 -52359.62 -16246.47 -46732.23        NA

          EEV       VEV       EVV        VVV
1  -63608.10 -63608.10 -63608.10 -63608.099
2  -61841.66 -34080.16 -61607.21 -30159.420
3  -55557.69 -30847.13 -55544.67 -23858.614
```

```
4  -55962.14 -23774.18 -54503.52 -16189.688
5  -54770.96 -23035.85 -52425.97 -15683.165
6  -51079.46 -22280.10 -50885.20  -8490.054
7  -50742.57 -17595.84 -49182.94  -6850.630
8  -49898.17 -17196.31 -47516.72         NA
9  -48929.90 -17200.97 -46019.87         NA
10 -51647.60 -14273.84 -45893.55         NA

Top 3 models based on the BIC criterion:
    VVV,7    VVI,10     VVV,6
-6850.630 -7790.461 -8490.054
```

Table 4.1 shows the number of datapoints assigned to each cluster by model based clustering.

Table 4.1: Number of datapoints in each cluster grouped by BIC.

1	2	3	4	5	6	7
369	1581	520	529	438	697	416

Unlike the methods in partitional clustering for which the number of clusters aligned with our categorical variable of interest, with model based clustering the number of clusters is much larger, and it would therefore not be expected for these to coincide. The first two principal components are plotted in Figure 4.14, grouped by the clusters from model based clustering.

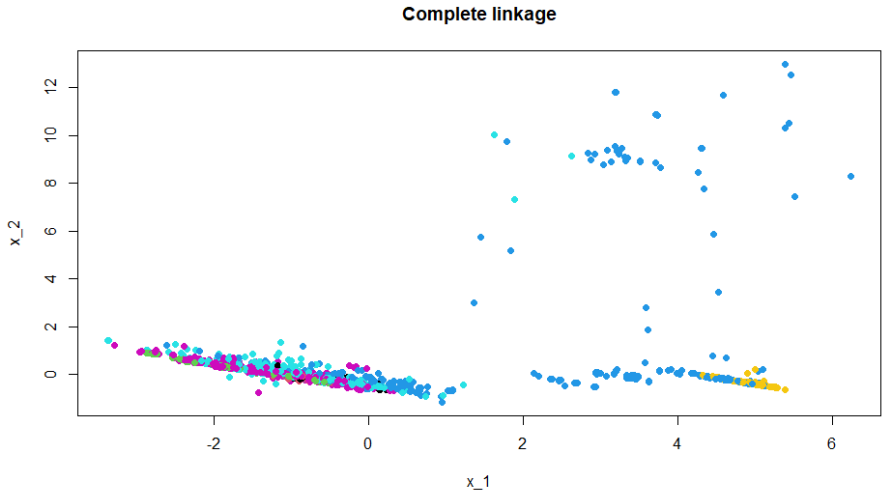


Figure 4.14: Scatter plots of the first two principal components colored by the seven groups by BIC.

5 Supervised Classification

In order to make a prediction of the categorical variable "Purpose", we made use of the supervised classification methods KNN, LDA, QDA, Naïve Bayes and Logistic regression. What makes a particular method better than another is determined by the nature of the dataset. Unlike unsupervised classification, these methods require that the data is split into two different samples, the training sample to estimate the parameters and the testing sample to measure how good the classifier is.

5.1 K-Nearest Neighbors

With the k-nearest neighbors (KNN) method, there are different ways of calculating k. Here we will be using Leave one out cross validation (LOOCV). After scaling the variables and estimating the error rate (LER), we can represent the LER for different values of k and select the one that returns the minimum value of the error rate (k=3 in this case), in Figure 5.1.

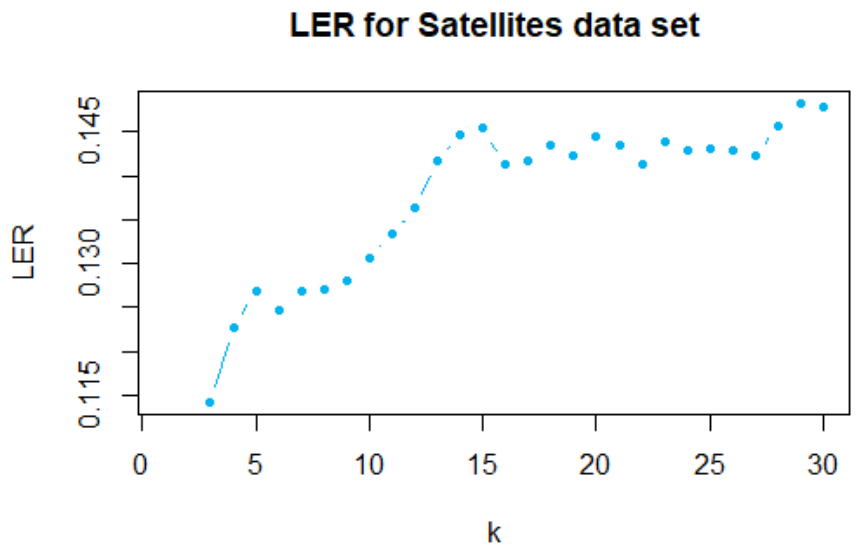


Figure 5.1: K vs LER

Using k=3, we were able to classify each observation in the test data. This gave us Table 5.1 which perfectly encapsulates the performance of this method:

Table 5.1: Confusion table of kNN.

BELONGS TO	CLASSIFY IN		
	Communications	**Earth observation**	**Others**
Communications	829	16	8
Earth observation	27	246	40
Others	22	58	119

This gives us a test error rate (TER) of 0.125 which tells us that 12.5% of the test data set was incorrectly classified.

5.2 Methods Based on the Bayes Theorem

There are three alternative methods to KNN that are based on Bayes Theorem where the idea is to obtain the conditional probabilities $Pr(y = k|x = x_0)$ where $k = 1, 2, 3$ since we have three classifications based on Purpose: Communications, Earth Observation and Others.

Making different assumptions regarding the set of probability density functions $f_1(\cdot|\theta_1)$, $f_2(\cdot|\theta_2)$, $f_3(\cdot|\theta_3)$ gives us the different methods we will use.

5.2.1 Linear Discriminant Analysis

The first method is called Linear Discriminant Analysis (LDA) where the assumption is made that the set of PDFs are multivariate Gaussian and share a common covariance matrix, Σ.

Using R, we estimated the unknown parameters with the training data set and used it to classify the observations in the test data set using the LDA method. This gave us Table 5.2 which shows that with this method 785, 166 and 72 satellites were correctly classified as having the purpose Communications, Earth Observation and Others, respectively.

Table 5.2: Confusion table of LDA.

BELONGS TO	CLASSIFY IN		
	Communications	**Earth observation**	**Others**
Communications	785	14	54
Earth observation	117	166	30
Others	70	57	72

This means we have a test error rate (TER) of 0.251.

5.2.2 Quadratic Discriminant Analysis

If instead we make the assumption that the set of PDFs do not share a common covariance matrix, Σ, but rather each PDF $f_k(\cdot|\theta_k)$ has a corresponding covariance matrix, Σ_k, we have our second method which is called Quadratic Discriminant Analysis (QDA).

Using R, we estimated the unknown parameters with the training data set and used it to classify the observations in the test data set using the QDA method. This gave us Table 5.3 which shows that, compared to the LDA method, more Earth Observation satellites were correctly classified but the opposite is true for satellites with classifications Communications and Others.

Table 5.3: Confusion table of QDA.

BELONGS TO	CLASSIFY IN		
	Communications	**Earth observation**	**Others**
Communications	748	94	11
Earth observation	61	234	18
Others	53	108	38

This gives us a test error rate (TER) of 0.253.

5.2.3 Naïve Bayes

The LDA and QDA methods necessitate the estimation of covariance matrices and the use of their inverses. This becomes more and more difficult as the number of predictors increases. The Naïve Bayes (NB) method gets around this by making the assumption that the predictors are independent variables, thus reducing the covariance matrices to 0.

Using R, we estimated the unknown parameters with the training data set and used it to classify the observations in the test data set using the NB method. This gave us Table 5.4 which shows that, compared to the QDA method, slightly more Earth Observation satellites were correctly classified but this was the case for much fewer satellites with classifications Communications and Others.

Table 5.4: Confusion table of NB.

BELONGS TO	CLASSIFY IN		
	Communications	**Earth observation**	**Others**
Communications	689	161	3
Earth observation	56	252	5
Others	67	124	8

This gives us a test error rate (TER) of 0.305.

5.3 Logistic Regression

Another good method for classification problems is logistic regression (LR). Typically, this method is used when the variable to be predicted is categorical and the outcome is binary, however, in our case we make use of multinomial logistic regression so that we can work with our three different classifications.

With R, we were able to use LR to classify each observation in the test data which gave us the Table 5.5.

Table 5.5: Confusion table of LR.

BELONGS TO	CLASSIFY IN		
	Communications	**Earth observation**	**Others**
Communications	784	40	29
Earth observation	104	182	27
Others	69	57	73

This gives us a test error rate (TER) of 0.239.

5.4 Comparison of Methods

Different approaches were attempted for each method in order to get lower values of TER, such as the elimination of uninteresting or highly correlated predictors (log_Apogee, log_Perigee, period), even performing the classification with single variables, however, none of these techniques were effective.

In order to sum up and make a comparison of the different methods, we can take a look at Table 5.6. It is a very good overall view, as we see how each approach differs from the real response variable in the test data set.

Table 5.6: Number of satellites classified in each group by method.

METHOD	**KNN**	**LDA**	**QDA**	**NB**	**LR**	**Real Y_test**
Communications	878	972	862	812	957	853
Earth Observation	320	237	436	537	279	313
Others	167	156	67	16	129	199

What 5.6 does not show is how many satellites were incorrectly classified, for instance the QDA method classified 862 satellites as Communications satellites which is very close to the true number (852) but it might be that many of these were incorrectly classified. Thus, to see method is best for our dataset, we really on the values we got for TER, in Table 5.7.

Table 5.7: TER values by method.

METHOD	**kNN**	**LDA**	**QDA**	**NB**	**LR**
TER	0.125	0.251	0.253	0.305	0.239

It seems that none of the Bayes theorem-based methods worked well enough for this particular dataset, in particular the NB method which we expected. Despite kNN being one of the simplest classifiers, the table above shows that it is the indisputable winner to predict the variable purpose in the satellites dataset, far ahead of the second best method, Logistic Regressor.