ETHNICITIES, ARRESTS AND UNEMPLOYMENT

AHPER TEAM

Minasian Levon, Bethany Dillon, Jonathan Clarke, Ilias Posidiz, Ibtihaj Naeem, Feidas Antoniou

PRACTICAL BUSINESS ANALYTICS COM3018 GROUP COUSEWORK

Contents

1. Project Definition

1.1 Background context:

In the recent years, conversations around racism and xenophobia have sparked much debate in the United Kingdom, becoming the catalyst for extremely polarising events such as the Brexit referendum and the election of Boris Johnson as prime minister, which have completely changed the country's political landscape and deeply affected the lives of millions of citizens in ways that are not quite quantifiable.

There seems to be this prenotion ingrained in the minds of many, especially the older generations in Britain, that the BAME population of the United Kingdom are responsible for the majority of crimes committed within the Kingdom. The idea is deeply rooted within the psyche of the British society; however, it is seldom supported with evidence strong enough to back these claims. Often, it appears to be propaganda rooted in post-colonial sentiments of nationalism and imperialism in a desperate attempt to re-establish the status quo of the British Empire.

1.2 Aims and Objectives:

BAME groups, and specifically young black males are overrepresented in all stages of the criminal justice system, yet only account for 8% of all arrests, whereas white individuals account for 79.6% of arrests - individuals of Asian, Mixed, Other or Unknown ethnicity making up the other 12.4% of arrests. However, the rate of arrest is hugely inflated for within ethnic minorities. The notion that peoples from a black ethnic background are committing more crime is archaic a logical non-sequitur, but there are other factors that may correlate to and explain these statistics.

A certain contributor to these statistics, and one that is frequently mentioned with regards to the rising crime rate, especially within ethnic minority groups, in the UK is the unemployment rate. This is a topic even more relevant now in the wake of COVID-19 as the unemployment rates in England and Wales have dramatically increased over the past few months, and with the country now in a second lockdown, are only bound to increase. With the already increased rates of unemployment within ethnic minorities, stemming from prejudice, a collapsing education system and lack of opportunities, a correlation between unemployment rates and arrest rates becomes evident within an ethnic context, and provides a plausible explanation for the grossly inflated rates of arrest within minority ethnic groups. With unemployment rates for ethnic minorities higher than for people from a White background, it would suggest that this could be, at least in some part, accountable for the inflated rate of arrests for minority ethnicities. As is the case with all models though, correlation does not necessarily lead to causation- something that an in depth understanding of the data being used will help us avoid assuming.

- The first and foremost aim of this project is to investigate whether a correlation between unemployment rates and arrest rates within different ethnicities as the one discussed above exists, and whether it can be demonstrated.
- The second aim of the project would be to provide local authorities, presiding government bodies and law enforcement agencies with the results of the analysis, granted the data yields satisfactory results and is able to outline some fashion of correlation, in the hope that:
    - Efforts are made to educate the public about the reality of the crime rates in the U.K, and whether BAME are more likely offenders compared to citizens of white background
    - Government policies can be adjusted to combat unemployment and crime rates simultaneously, by providing more individuals from BAME groups access to opportunities

1.2.1 Hypothesis:

The hypothesis of our project is therefore based on a simple premise: do unemployment rates affect the crime rates, and therefore the arrest rates within different ethnic groups, with a focus on minorities?

- For the hypothesis to prove correct, the results from the analysis should discern that:
    - For a given ethnic group, an increase in the unemployment rate over a period of time should result in an increase on the arrest rates within that same period of time.
- And vice versa:
    - For a given ethnic group, a decrease in the unemployment rate over a period of time should result in a decrease on the arrest rates within that same period of time.

2. Project Management

Below, we are discussing the techniques employed by the group, as we took a holistic approach and decided to tackle the project by having two teams working in parallel and concurrently, while also having general recurring group meetings to share progress and discuss the next steps. Since this was a group project, the two main concerns were the management of the team members and the fair and equal distribution of responsibilities and workload, as well as the management of the timelines to guarantee that all deliverables are ready by the deadlines.

2.1 Team Management:

The first step during the planning process for our project work was determining how the team would be managed, so that each member would be able to make a significant contribution to the project, but also to ensure that the workload was divided equally and fairly amongst the team.

2.1.1 Sub-Teams:

The team came to the unanimous consensus that the best way for us to handle the project would be by forming sub-teams, each with well-identified and clear tasks to accomplish.

- Team A, which we referred to as the Programming Team, dealt with the technical implementation of the model, which was the most challenging aspect. The team had three members and they have 2 weekly meeting where the member would share their progress and work through the most challenging tasks together.
- Team B, also referred to as the Documentation Team, had the responsibility of writing up the documentation for the project, which was believed to be the most time-consuming aspect. Although there is a dependency between the implementation and the and the documentation, the team managed to implement a method that allowed parallel tasks being completed between Team A and Team B. This team also consisted of three members, and they team would have 1 weekly meeting to share progress, review the documentation and discuss the further steps.

2.1.2 Team Meetings:

Another technique employed for the team management was having regular team meetings and check-ins. The team scheduled 2 weekly team meetings, generally on Monday and Thursday, where all members of the team would join to discuss the overall progress of the

project for each team. During these meetings, the code team would also pass new or any additional information over to the documentation team, and the plan for the upcoming tasks were discussed in detail, so that each team would have tasks to complete by the next meeting. The meetings were an efficient way to streamline the process given that due to COVID it was not possible for the team to meet physically and share each other's work, so by using online shared repositories and documents both teams were able to have their members share their work and collaborate in real time.

2.2 Workload Timeline:

The initial timeline for the project was outlined in the project plan document, where important submission deadlines as well as internal deadlines were identified, and tasks were outlined to be completed accordingly.

There were changes to the original workplan, as the time needed for the implementation was underestimated and more time was required by the coding team to complete the development of the model.

Fortunately, this did not have repercussions on the submission of the project, as the deadline of the project was shifted by two weeks, which gave both teams enough time to complete all outlined tasks to a high quality.

3. Data Exploration & Preparation (Pre-Processing):

The project makes use of the CRIPS-DM methodology which is a common approach to data mining. Before we could start the model development for this project, the data needed to be explored, pre-processed, and prepared. In this way, the data would then both be easier to be understand and cleaner for the model to work on.

During the first stage, we familiarised ourselves with the data and encoded the variables of Region, Year and Ethnicity from the dataset of "Number of Arrests" and the variable of Ethnicity from the dataset "Unemployment by Region". Since we are trying to find correlations between ethnicities in two datasets, we first exposed the values of each dataset for their respective "Ethnicity" field and found that one dataset had more unique values than the other. For example, we identified, that in one of the datasets there were many ethnicities sub-groups. For example, for the Asian ethnicity, there were subgroups Asian and Asian Other. Initially these were not combined, but we later recognised that Asian Other was a subset of Asian, and the two were therefore merged.

```
Browse[2]> unique(arrests_dataset$Ethnicity)
 [1] "Asian"                                  "Black"
 [3] "Mixed"                                  "Other"
 [5] "Unreported"                             "White"
 [7] "Any other asian"                        "Any other black background"
 [9] "Any other ethnic group"                 "Any other mixed/multiple ethnic background"
[11] "Any other white background"             "Bangladeshi"
[13] "Black African"                          "Black Caribbean"
[15] "Chinese"                                "Indian"
[17] "Mixed White and Black Caribbean"        "Pakistani"
[19] "White British"                          "White Irish"
[21] "All"                                    "Mixed White and Black African"
[23] "Mixed White and Asian"
Browse[2]> unique(unemployment_dataset$Ethnicity)
 [1] "All"                       "Asian"                 "Asian Other"           "Black"
 [5] "Indian"                    "Mixed"                 "Other"                 "Other than White"
 [9] "Pakistani and Bangladeshi" "Unknown"               "White"                 "White British"
[13] "White Other"
```

Figure 3.1: Field uniqueness – before

Thus, the arrests dataset's Ethnicity field were reduced to match the values from the field of the unemployment dataset. This was done by collecting the sub-groups of different ethnicities into the main ethnicity, so that each dataset has the same values for the field, as discussed in the example above. After the merging, the field is reduced to five values, which correspond to the to the five ethnicities we are interested in.

```
Browse[2]> unique(arrests_dataset$Ethnicity)
[1] "Asian" "Black" "Mixed" "Other" "White"
Browse[2]> unique(unemployment_dataset$Ethnicity)
[1] "Asian" "Black" "Mixed" "Other" "White"
```

Figure 3.2: Field uniqueness – after

Similar problems were also encountered in the region/geography field and the time (year) field, which respectively recorded where the data was observed, and when the data was recorded. These two fields went through the same process as that described above for the

ethnicities field. Smaller regions from the UK were merged into the major regions of the country, while data recorded from with different year formatting but that occurred within the same year were all merged into the same year.

In the dataset for “Number of Arrests”, the geographical location used corresponded to cities, while in the “Unemployment by Region” dataset, the geographical locations values corresponded to regions (e.g., East of England). As a result, the cities from the dataset “Number of Arrests” had to be merged into regions. Once these were merged, the numerous rows in the dataset of “Number of Arrests” were reduced. For instance, if previously two cities fitted into the same region, now they were merged together, there would only be one row for the region.

For the time field, in the dataset “Number of Arrests” the years were in a financial year format: i.e., 2007/2008, while in the dataset of "Unemployment by Region” , the years were represented as calendar years, e.g., 2007. So, the dataset of “Number of Arrests” needed to be changed again to be the same as the “Unemployment” one.

Furthermore, once all the fields were standardised, we were able to continue exploring the datasets and tried to allocate a type to each field, but we have found that due to certain inconsistencies, such as “missing” or “n/a values”, certain fields were registering as Symbolic when in fact they were Numeric. For example, in the number of arrests field in the “Number of Arrests” dataset, when there was no data for the number of arrests, the dataset had this field as either a “-“ or “N/A”. To combat this, the pre-processing stage also included a set of instructions to remove rows or modify them. By modify, we mean to replace the unknown sample by the average of similar samples.

This helped clean the datasets by reducing the amount the data the models would have to process by removing redundant information. It also allowed for all the fields that the data type models required to be numeric.

field	types
Measure	SYMBOLIC
Measuretype	SYMBOLIC
Ethnicity	SYMBOLIC
Ethnicitytype	SYMBOLIC
Time	ORDINAL
TimeType	SYMBOLIC
Region	SYMBOLIC
Age	SYMBOLIC
AgeType	SYMBOLIC
Sex	SYMBOLIC
Value	ORDINAL
confidenceinterval	ORDINAL
Numerator	ORDINAL
denominator	ORDINAL
sampsize	ORDINAL

Figure 3: Unemployment dataset

field	types
Time	ORDINAL
Region	SYMBOLIC
Ethnicity	SYMBOLIC
Numberofarrests	ORDINAL
Population	ORDINAL

Figure 4: Arrests dataset

Removing such fields also allowed us to perform field analysis. Here we could see the most common value for each field, and for Numerical fields their mathematical attributes. Looking at these figures also unveiled which columns were non-unique. We could see fields with 100% in their 'name' category indicating that there was variation in their data. These fields were thus noted to not be used as they would not affect the accuracy of the model.

	Field	Catagorical	Symbols	Name	Min	Mean	Max	Skew
5	Time	✖ No	-	0	2,006.00	2,011.50	2,017.00	0.00
11	Value	✖ No	-	0	3.00	10.18	29.20	0.75
12	confidenceinterval	✖ No	-	0	0.10	4.22	13.20	0.51
13	Numerator	✖ No	-	0	300.00	73,124.61	1,971,400.00	6.62
14	denominator	✖ No	-	0	6,600.00	1,192,596.34	28,423,100.00	6.24
15	sampsize	✖ No	-	0	101.00	9,718.48	257,987.00	6.27
1	Measure	✔ Yes	1	Unemployed(100%)	-	-	-	-
2	Measuretype	✔ Yes	1	Percentage of individuals unemployed(100%)	-	-	-	-
3	Ethnicity	✔ Yes	5	Asian(20%)	-	-	-	-
4	Ethnicitytype	✔ Yes	1	ONS 2011 5+1(100%)	-	-	-	-
6	TimeType	✔ Yes	1	year(100%)	-	-	-	-
7	Region	✔ Yes	12	All(8%)	-	-	-	-
8	Age	✔ Yes	1	All(100%)	-	-	-	-
9	AgeType	✔ Yes	1	16+(100%)	-	-	-	-
10	Sex	✔ Yes	1	All(100%)	-	-	-	-

Figure 5: Unemployment dataset field analysis

	Field	Catagorical	Symbols	Name	Min	Mean	Max	Skew
1	Time	✖ No	-	0	2,009.00	2,013.00	2,017.00	0.00
4	Numberofarrests	✖ No	-	0	0.10	2.36	8.70	1.62
5	Population	✖ No	-	0	4,106.00	1,937,079.43	48,209,395.00	6.51
2	Region	✔ Yes	12	South West(8%)	-	-	-	-
3	Ethnicity	✔ Yes	5	Asian(20%)	-	-	-	-

Figure 6: Arrests dataset field analysis

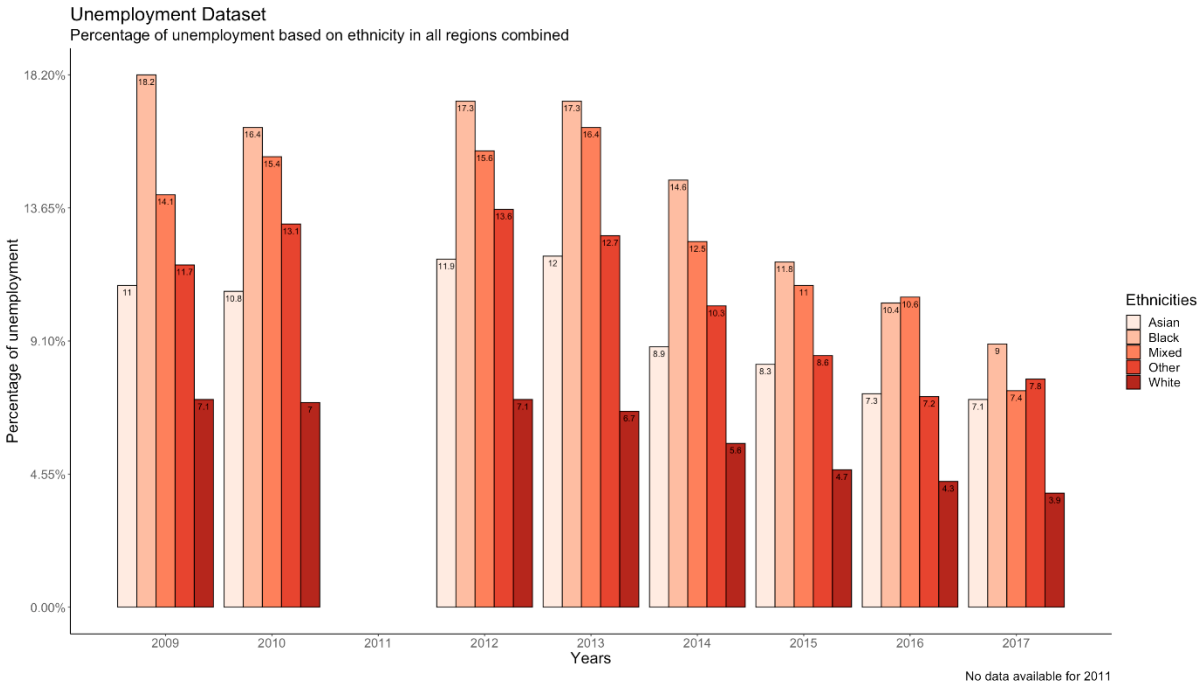


Figure 7: Unemployment dataset field analysis

This graph helps us to see the percentages of unemployment over the last few years based on the main five ethnicities in all regions combined. According to the graph, we clearly see that the percentages of unemployment is at its highest in 2009, and has decreased for all ethnicities by 2014, after which percentage of unemployment decreases every year for all ethnicities. The difference in increase/decrease for unemployment during 2009-2013 can be explained by the economic crisis which took place during this time.

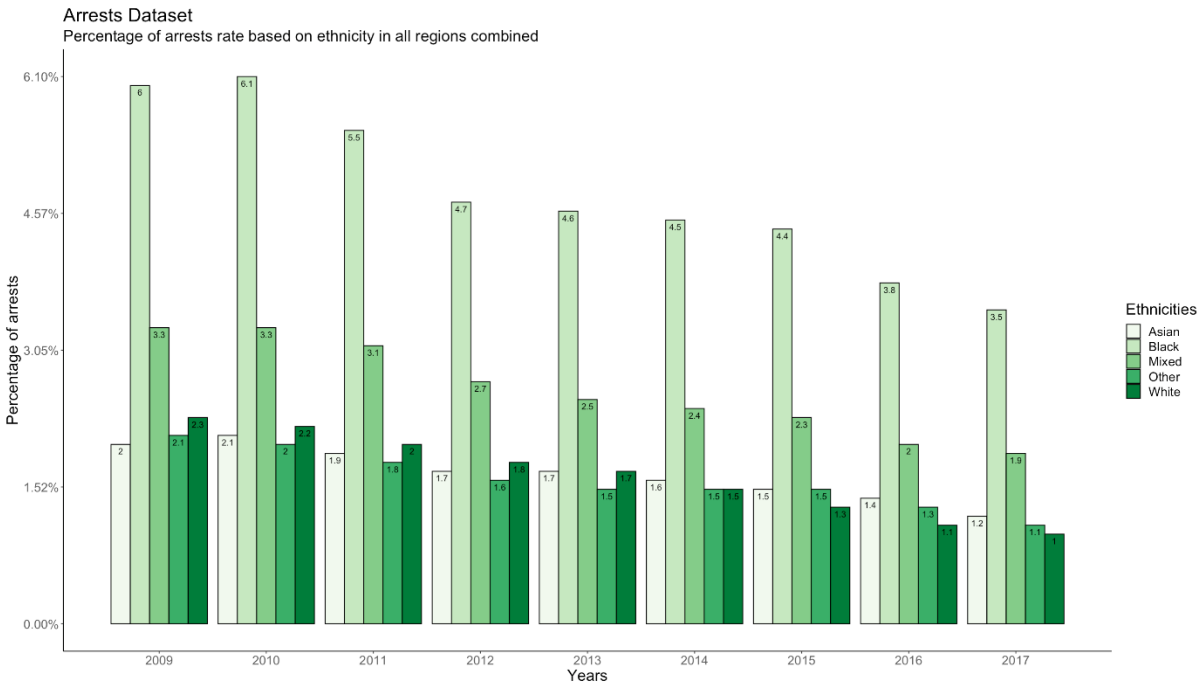


Figure 8: Arrests dataset field analysis

In this graph now, we can see the percentages of number of arrests over the last years that are based on the five main ethnicities in all regions combined. Firstly, according to the graph we clearly see that in 2009 and 2010 we have the highest number of arrests for all our ethnicities. Like the previous dataset of “Unemployment”, the percentages of numbers of arrests from 2010 and after, are continue gradually decreasing over the years for all the ethnicities. Lastly, we identify and notice that in the final year (2017), the percentages of number of arrests are at their lowest point within the graphed years.

According to the graphs of Unemployment and Number of arrests, based on ethnicities, we have concluded that our hypothesis was indeed correct! The percentages of unemployment over the years were decreasing, so too was the percentages of number of arrests for the same years. This brings us back to our hypothesis that confirms that when Unemployment is decreasing, the number of arrests is decreasing respectively as well.

4. Modelling

The processes discussed after this point have reduced the number of rows in each dataset so that both datasets have the same number of rows. This allowed us to merge the two datasets, creating a new 3rd dataset. The initial step to merging was reordering both datasets to be in the same order. Since we had used data tables, a joint function based on 3 conditions helped us to merge the 2 datasets. Those 3 conditions were the following: Time, Region, and Ethnicity. This allowed us to insert the correct unemployment value with the corresponding arrest value.

Figure 9 shows the arrests rate plotted against population for all ethnicities. We can see that there is no distinct correlation. The graph has 'chunks' of points, from each ethnicity respectively, which each suggest a vague line of best fit.

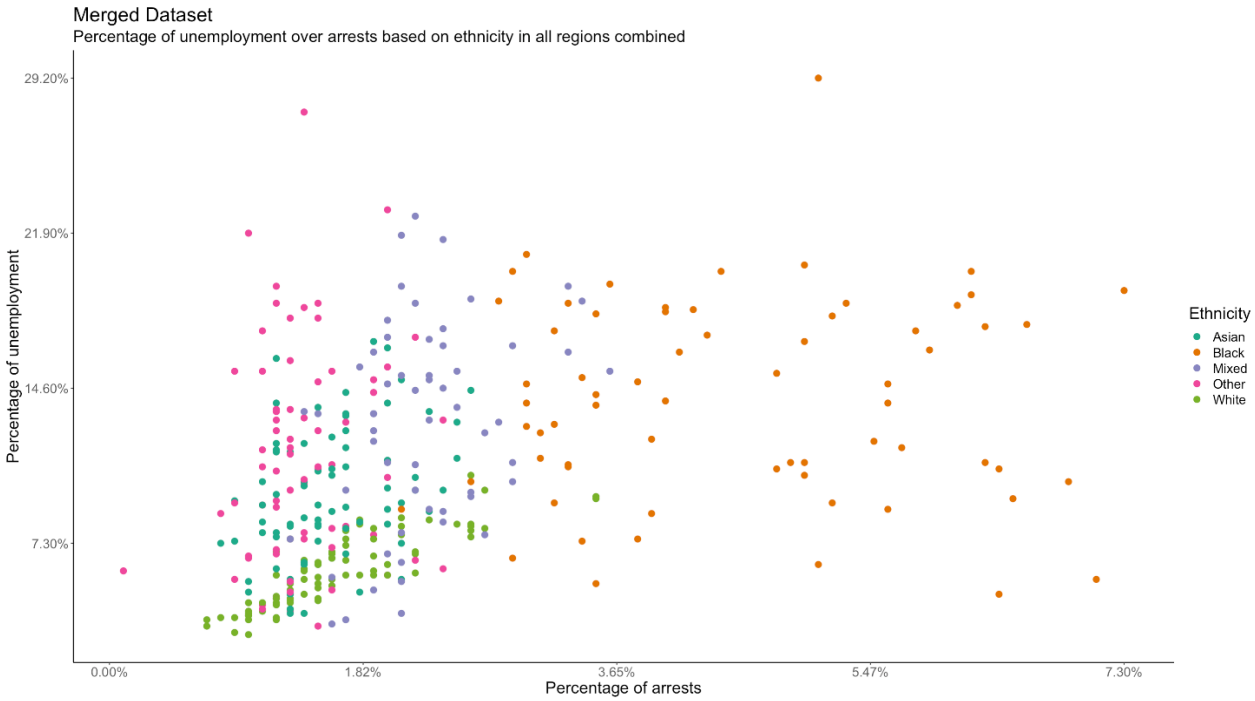


Figure 9: All ethnicities and all regions

In order to fully test our hypothesis, we split the dataset into ethnicities to train a model for each. This would allow us to meet our project objectives as we will be able to use the corresponding model when predicting figures for a given ethnicity.

After splitting the ethnicities, the models of which are described in further detail below, removed the 'All' Region entries, as all the other regions were a derivative of this parameter.

4.1 Individual Ethnicity Models:

There were two types of Supervised Learning approaches available to use when we were beginning to model our data. Regression and Classification. A Regression model is usually associated with a continues output value label, a numeric. Discreet outputs, where we label such as “good” or “bad” are used for classification. Therefore, since our project is trying to generalise and correlate arrests and unemployment data, we developed regression models. In Supervised Learning, we also have numerous algorithms we can use to model our data: Linear and Polynomial regression.

Linear Regression is a straight-line equation. It assumes the following equation.

$$y \approx \beta_0 + \beta_1 x + \varepsilon$$

We want to generalise our data and find a correlation essentially between arrests and unemployment and Linear Regression will produce a straight line of best fit.

Although splitting the models did create some improvements in accuracy, there were still some issues. We initially had split the data into a ‘train’ and ‘validation’ dataset, however due to the small number of samples and the large variation of population size between regions there was a high chance of underrepresentation. To combat the population difference first, we divided the total number of arrests by the population of a region. This allowed the model to represent each region fairly and so the correlation between unemployment and arrests improved for all ethnicities.

However, during data preparation a large amount of data had been removed from the dataset (ethnicity subgroups were discarded and value from cities in the same regions were summed together) and so during training, there was not enough data from each region so represent the true correlation within the data. Consequently, it was possible that the training records could have been filled with regions with high arrests and the validation set was filled with regions with low arrest, causing the linear model to be compromised. Our solution to this was to switch to K-Fold Cross-Validation to ensure that all regions were represented in the model. After experimenting with different values for k, we decided to use a value of 5 as it gave us an overall best result for our models.

In order to have another model to compare and contrast to, we also trained a polynomial regression model.

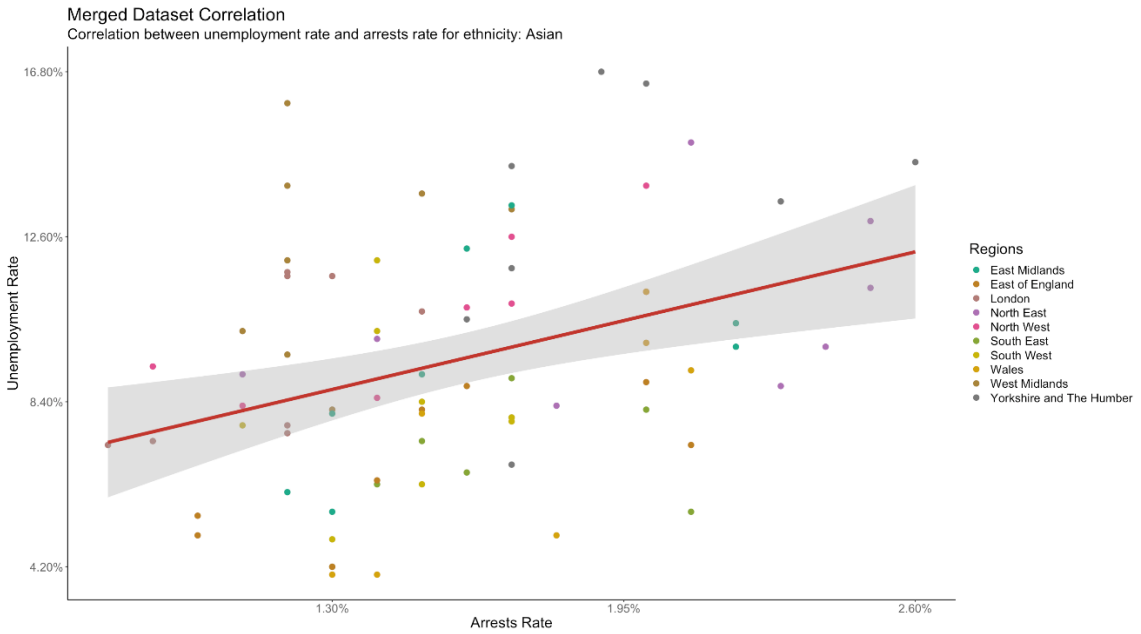
$$y \approx \beta_0 + \beta_1 x + \beta_2 x_1^2 + \cdots + \beta_n x_1^n$$

This model allows for non-linear relationships and can provide a better approximation of the relationship between the dependent and independent variable.

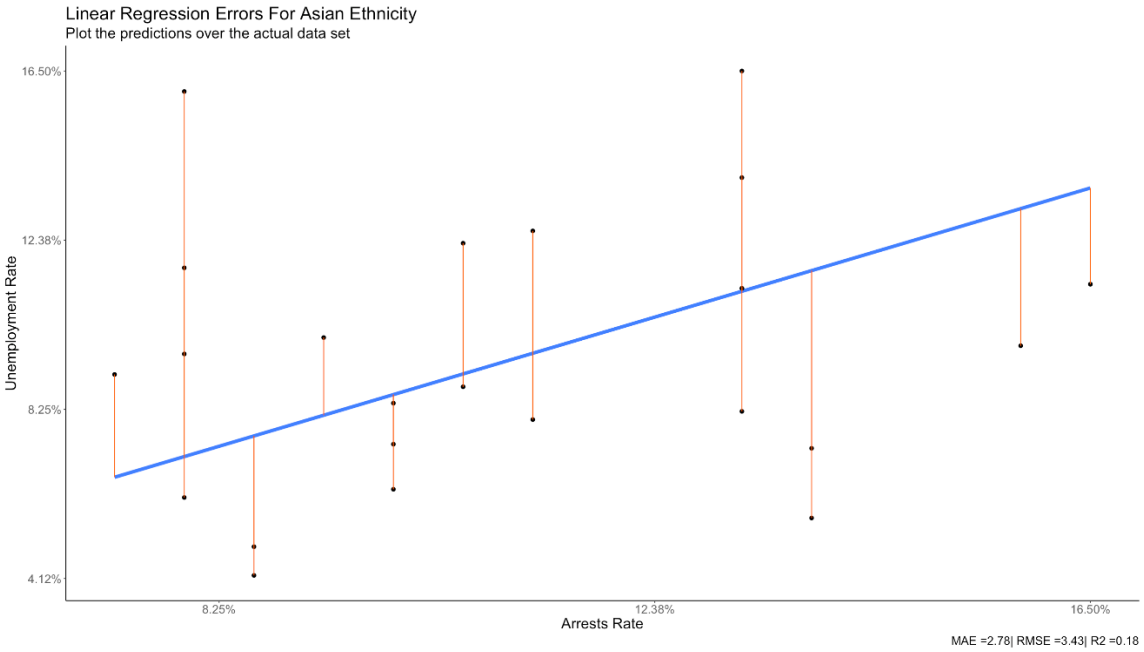
4.1.1 Asian Ethnicity

Firstly, according to the graphs below for Asian Ethnicity, we can clearly see the correlation between the unemployment and arrests rate. Our hypothesis confirms again that when the Unemployment rate is increasing, then the Number of Arrests is increasing, respectively.

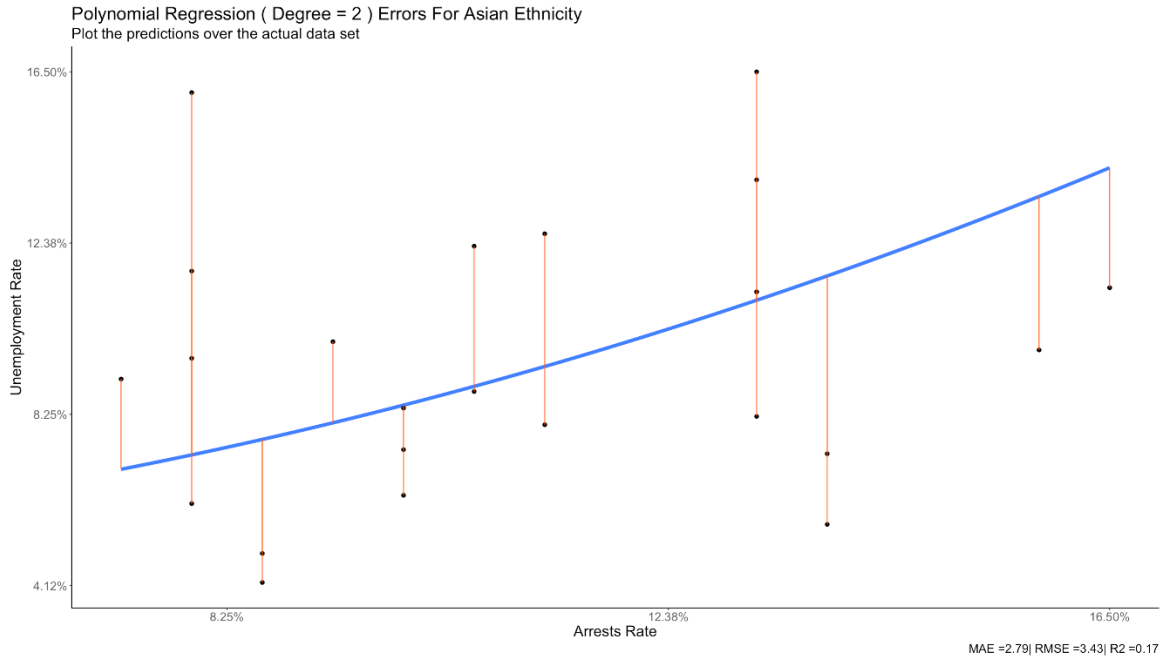
Correlation



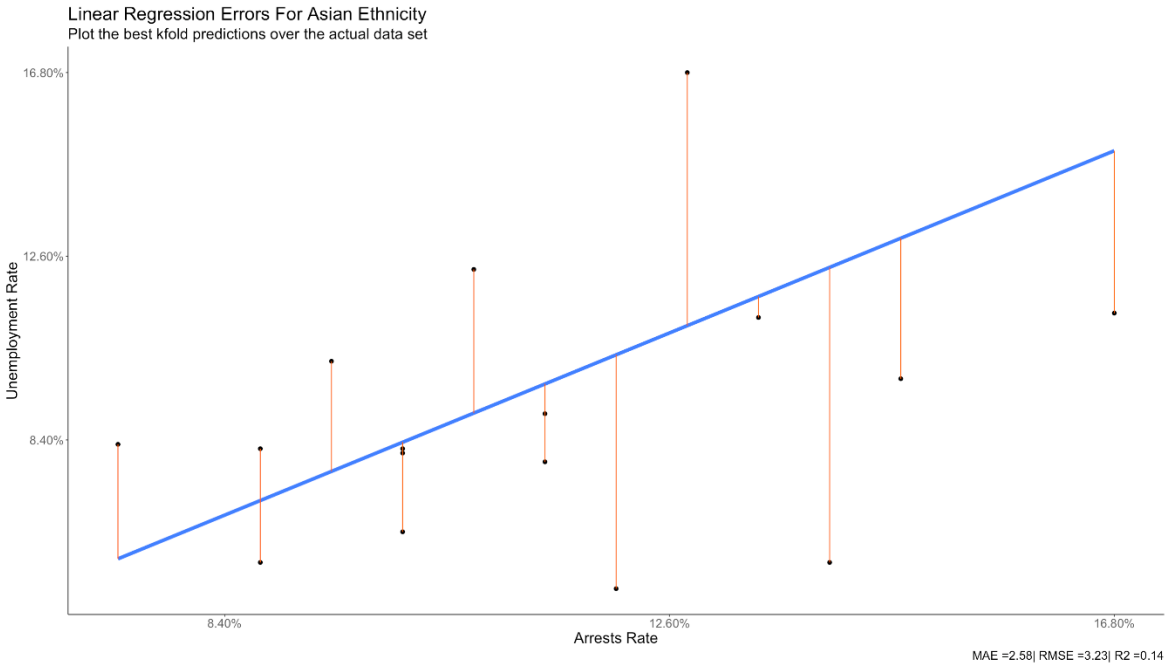
Linear using holdout



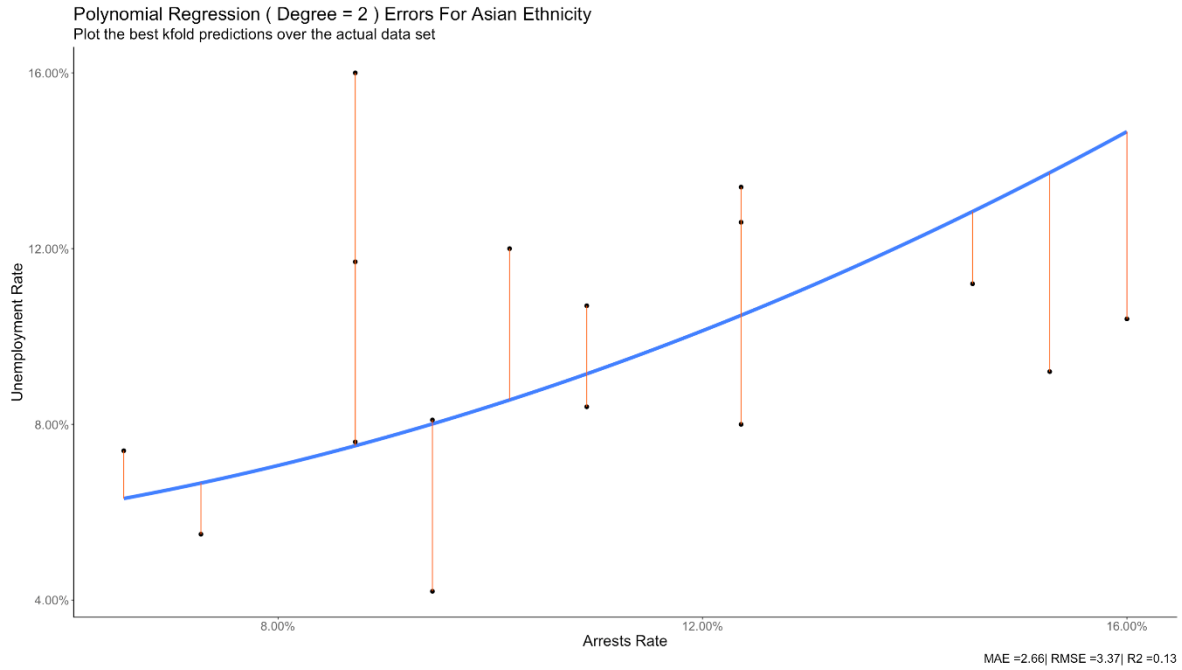
Polynomial using holdout



Linear using KFold



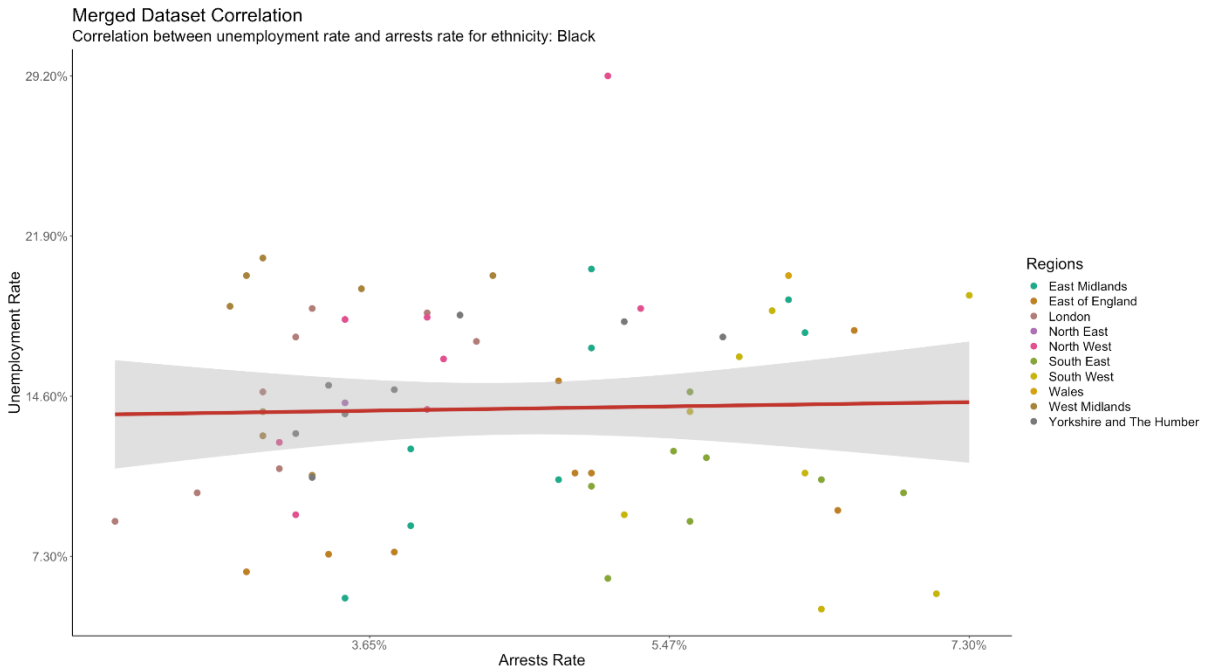
Polynomial using KFold



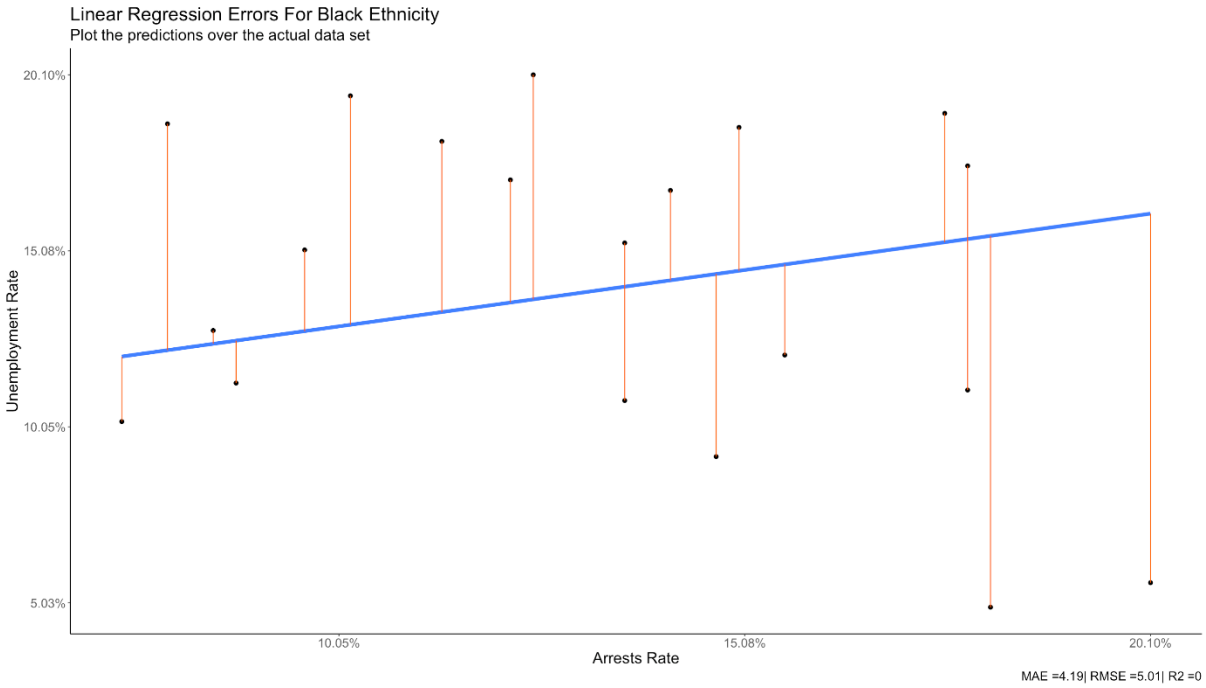
4.2.2 Black Ethnicity

The graph below plots unemployment rate against arrests rate for Black ethnicity. By looking at the points there is not a clear correlation between the unemployment rate and arrests rate. There is a correlation for each region, but we want to include all the regions. We also plot the predictions of errors in Linear and Polynomial Regression as well.

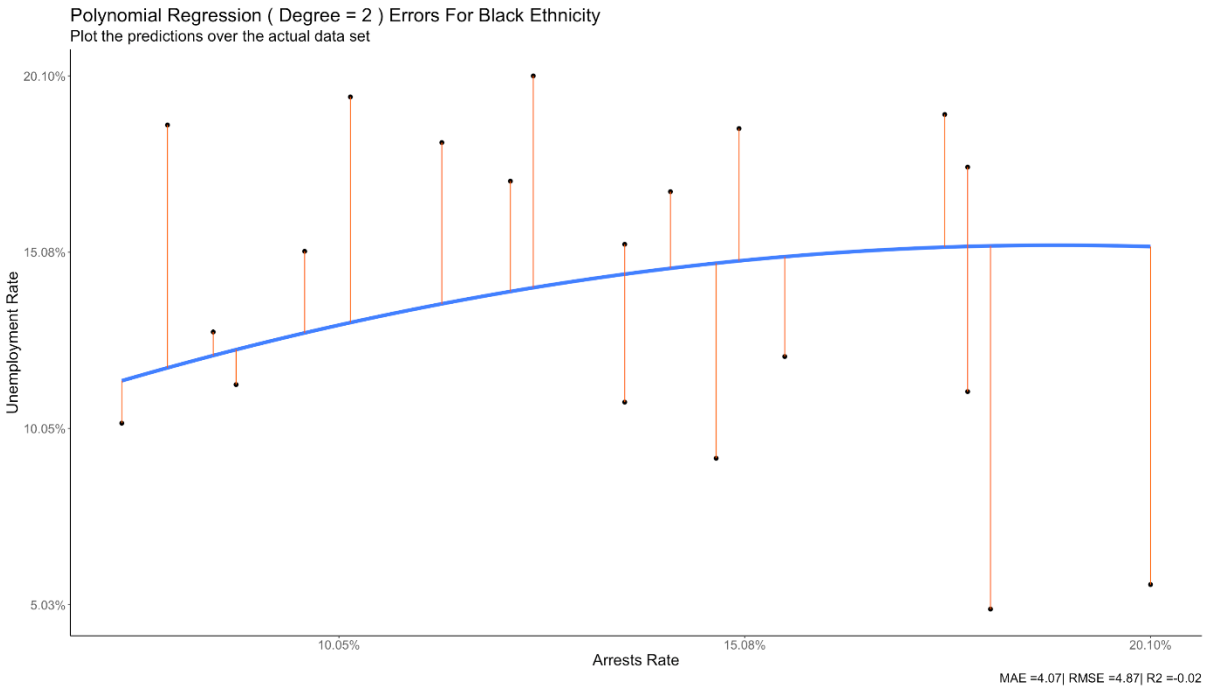
Correlation



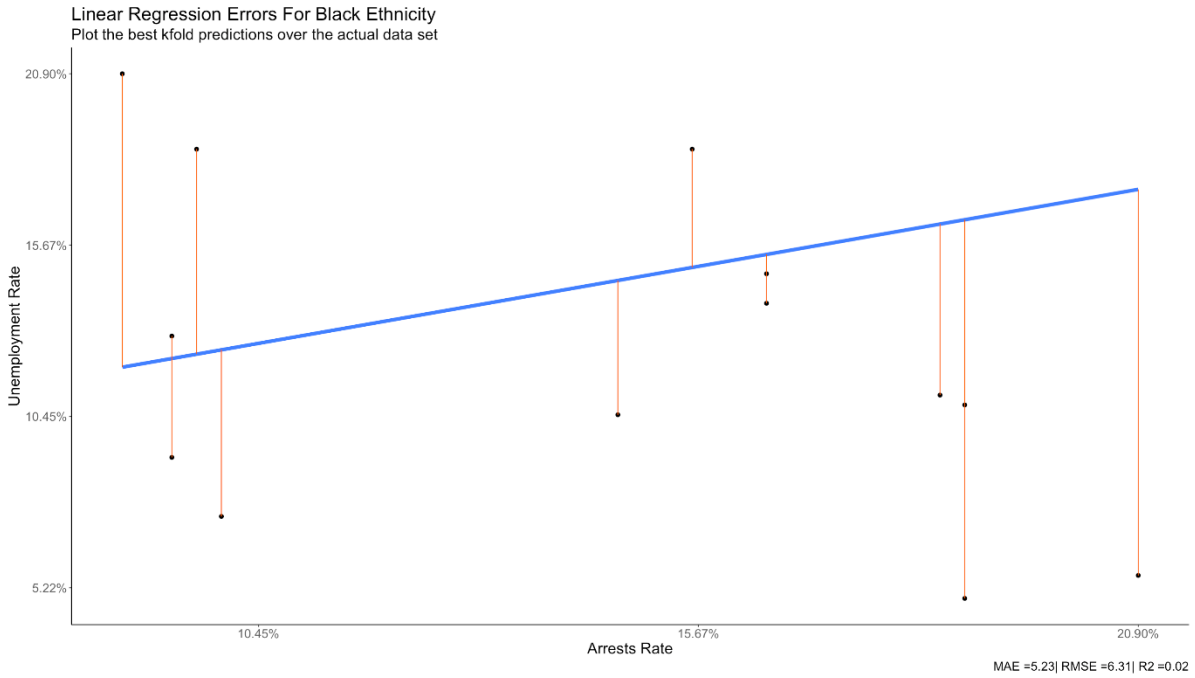
Linear using holdout



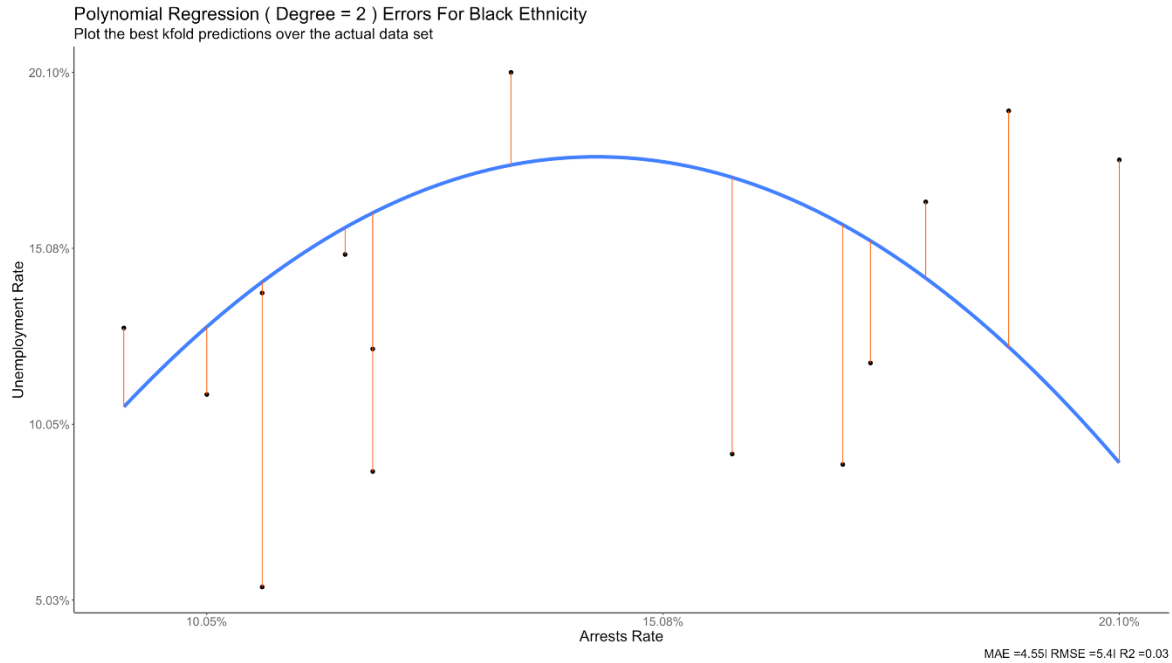
Polynomial using Holdout



Linear using KFold



Polynomial using KFold

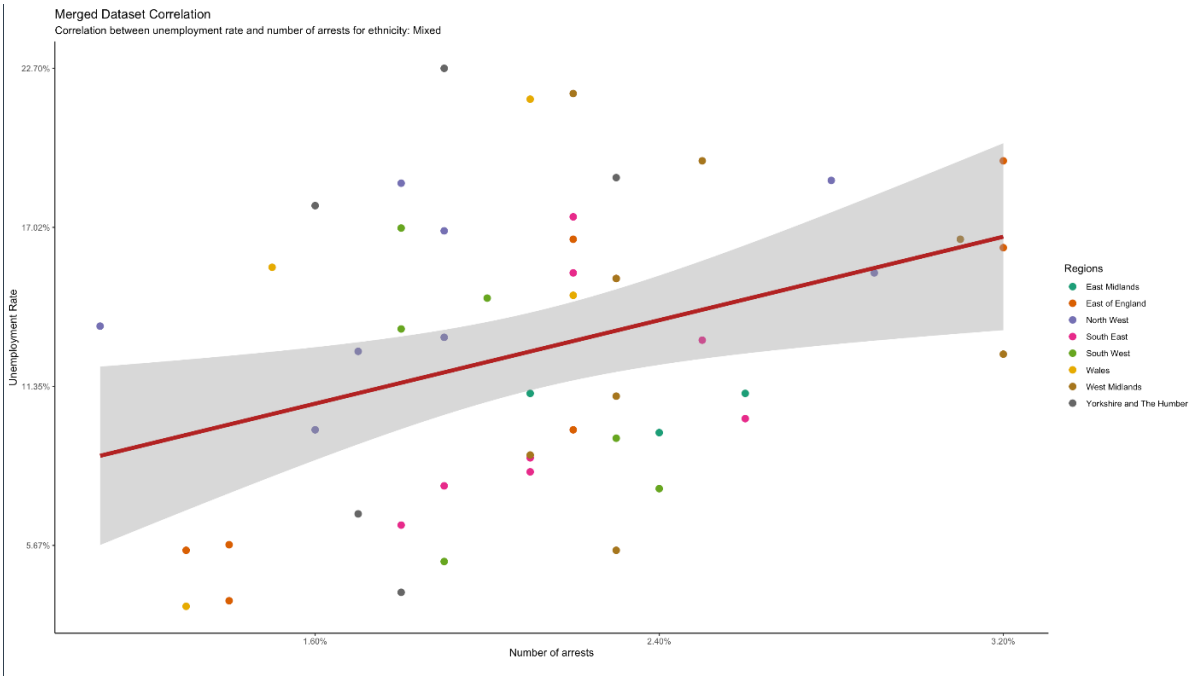


4.2.3 Mixed Ethnicity

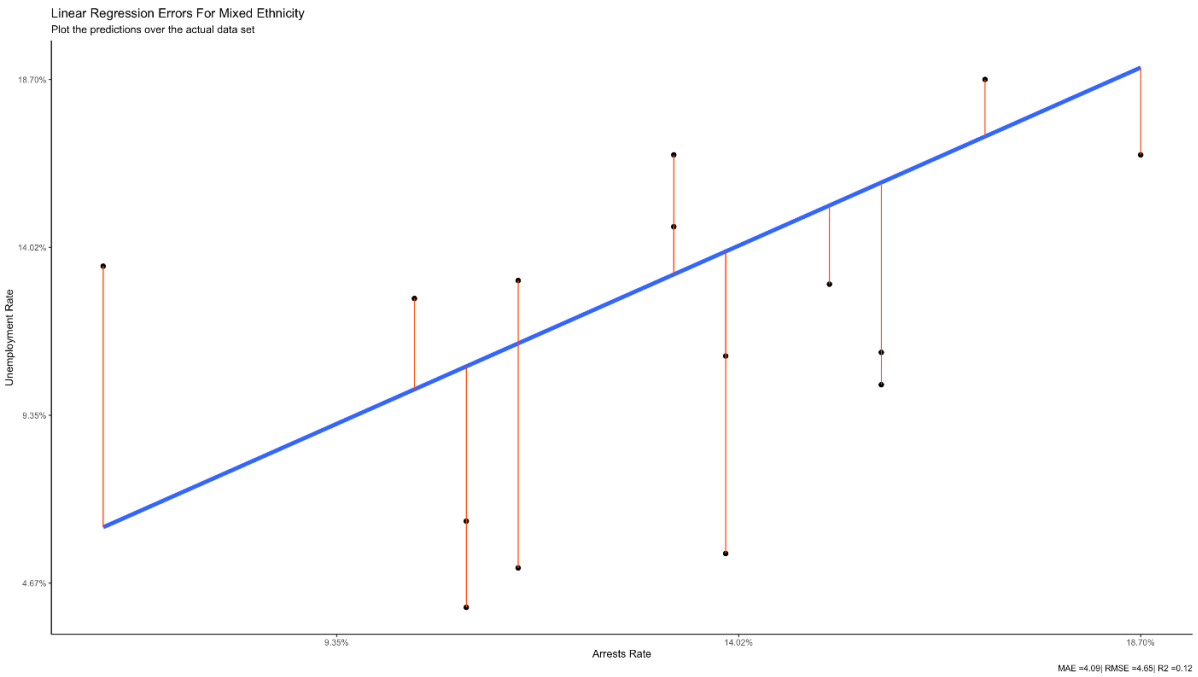
The graph below plots unemployment rate against the number of arrests for Mixed ethnicity. By looking at the points we can see a weak positive correlation between the two- as unemployment rate increases, so do the number of arrests, confirming our hypothesis. However, there are many outliers due to the variation in population density in different regions, the most out of any ethnicity. Even though our model uses a percentage for the

number arrests to mitigate the effects of population, the lack of enough sample data for Mixed ethnicity causes there to be several outliers.

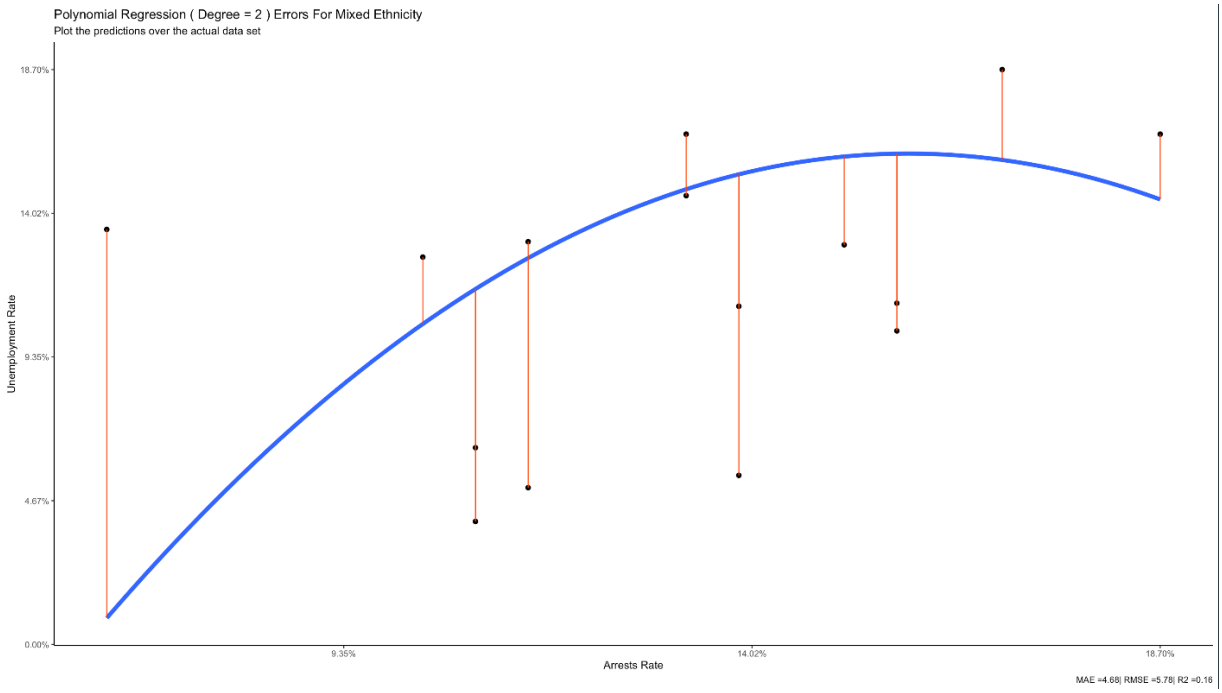
Correlation



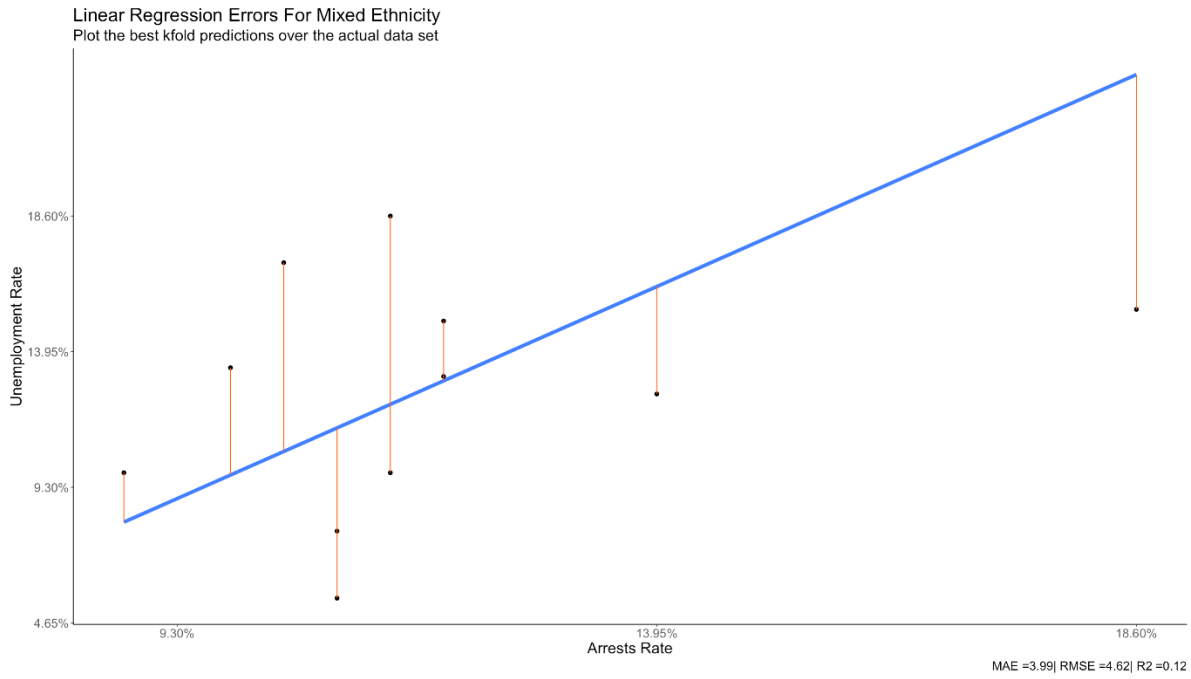
Linear using holdout



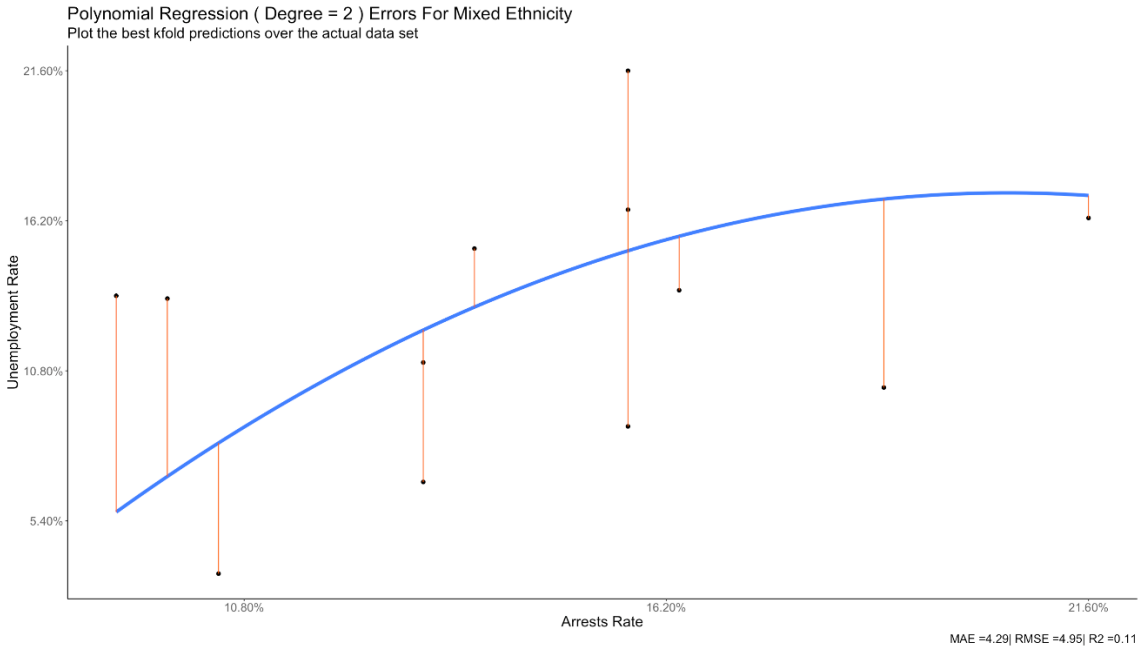
Polynomial using Holdout



Linear using KFold

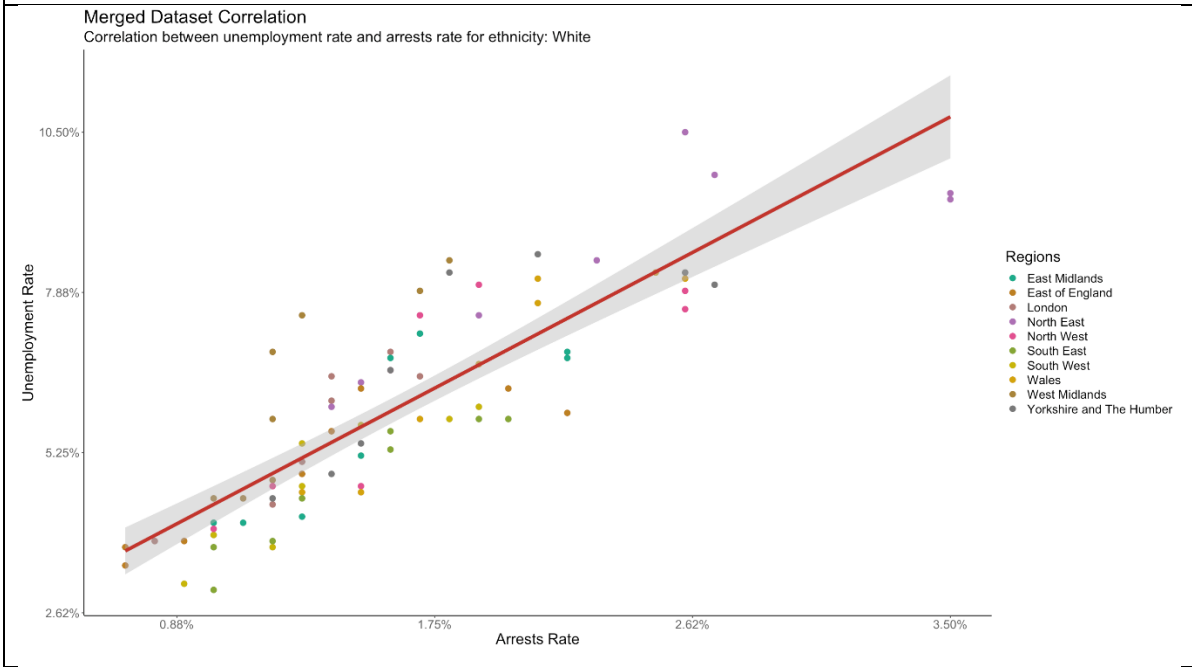


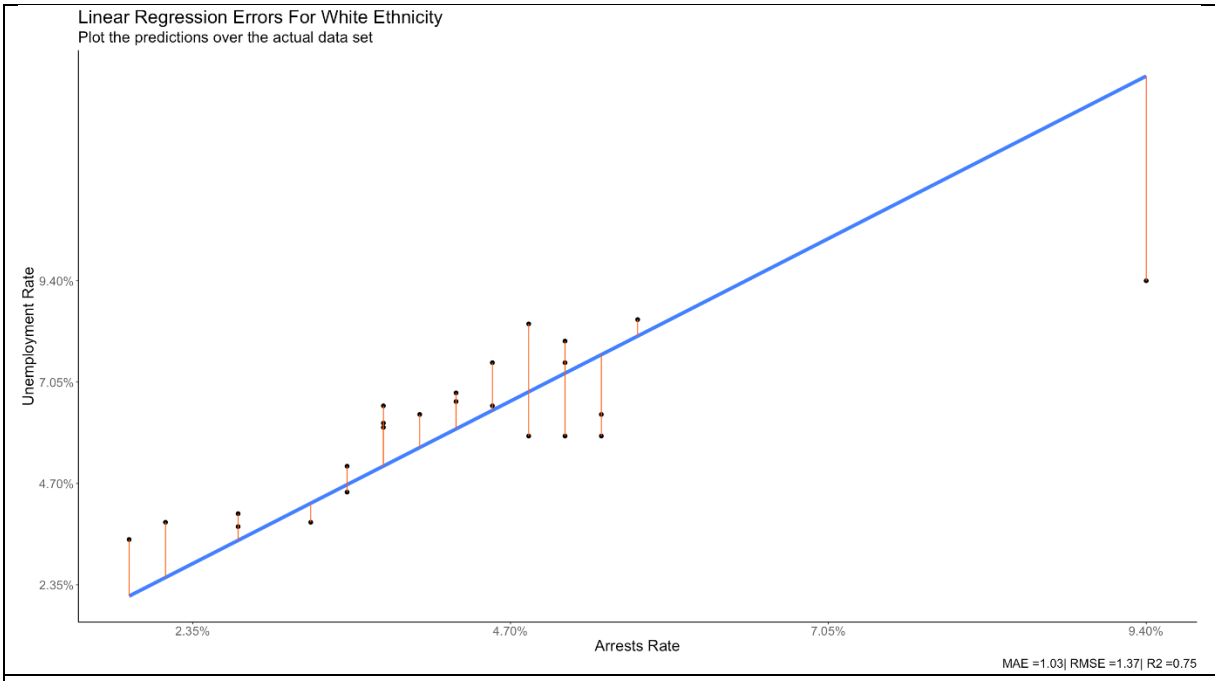
Polynomial using KFold



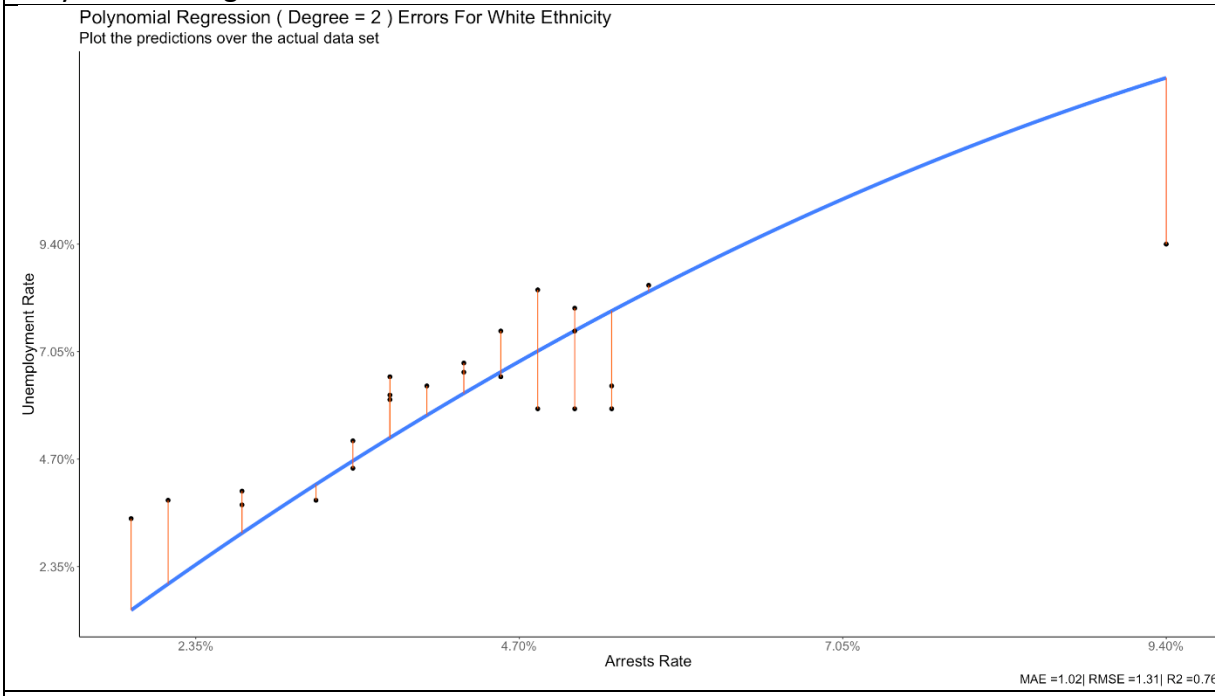
4.2.4 White Ethnicity

The graph below plots unemployment rate against the number of arrests for White ethnicity. By looking at the points we can see a string positive correlation between the two. As unemployment rate increases, so does the number of arrests, confirming our hypothesis. Due to the more uniform population density for this ethnicity, there is a smaller amount of outlier points.

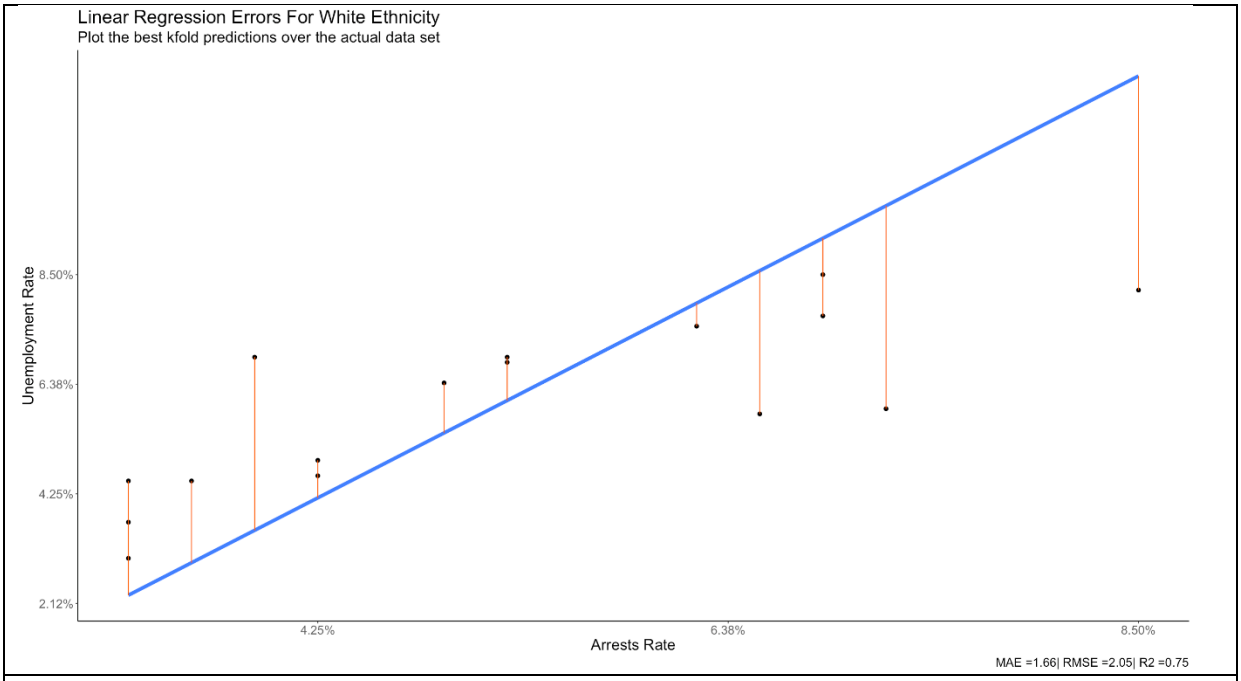
Correlation

Linear using holdout



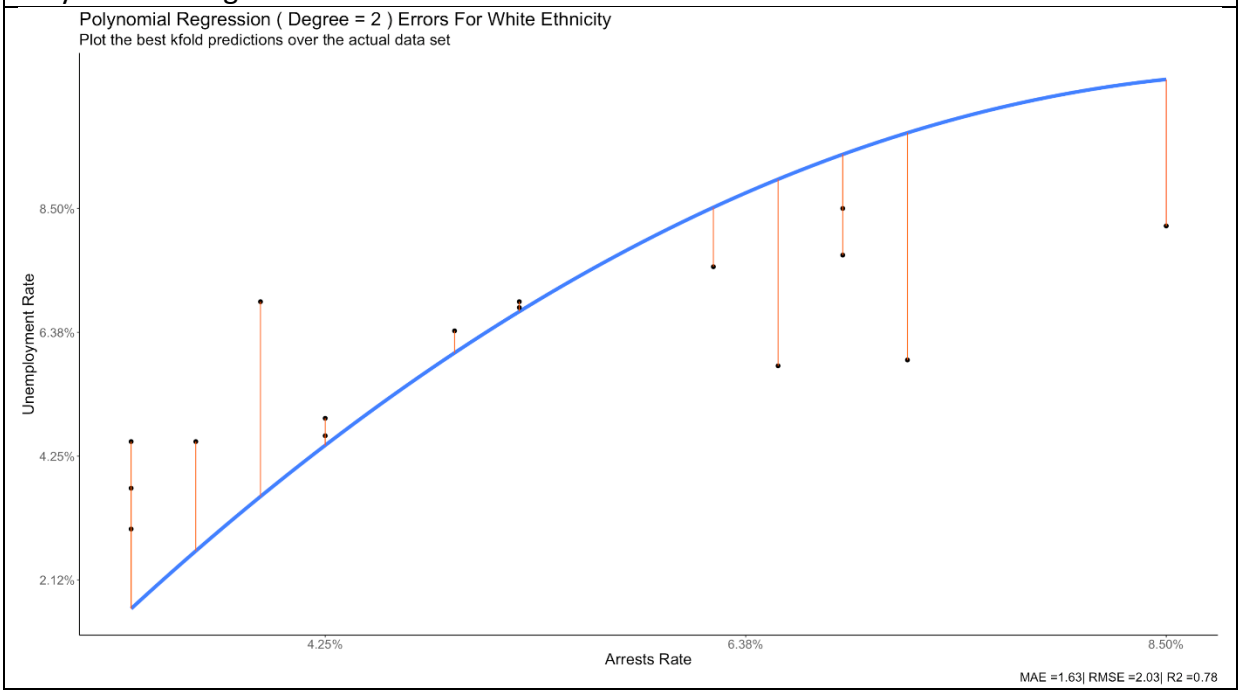
Polynomial using Holdout



Linear using KFold



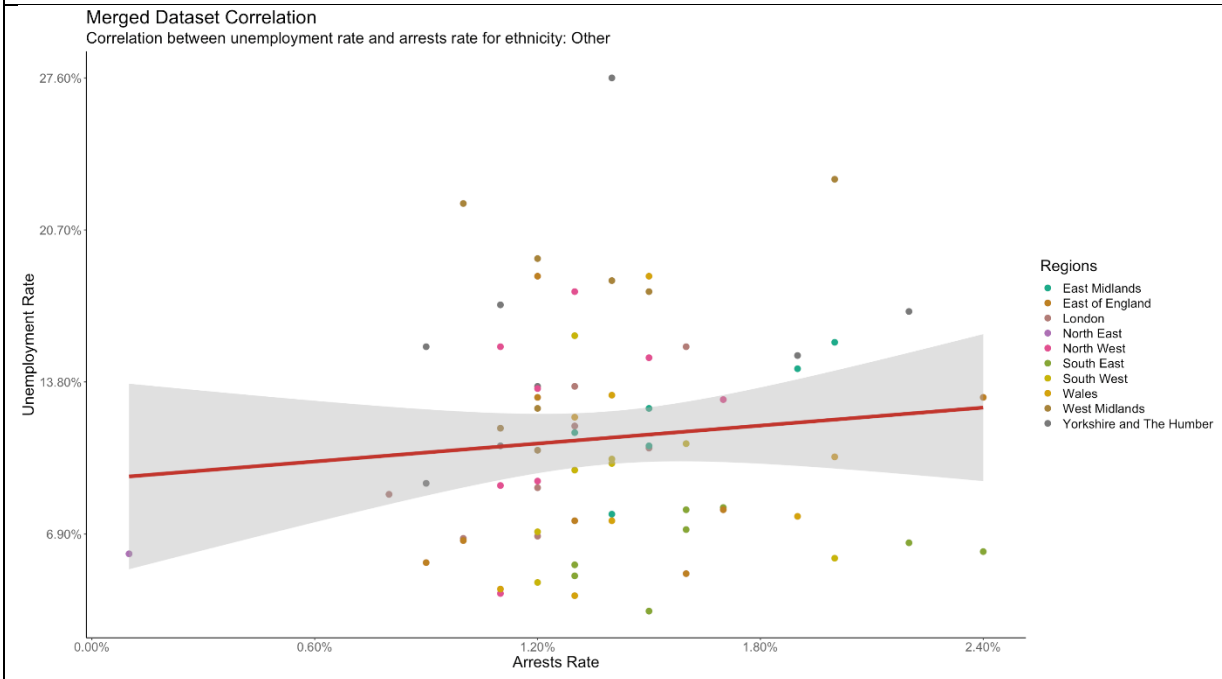
Polynomial using KFold



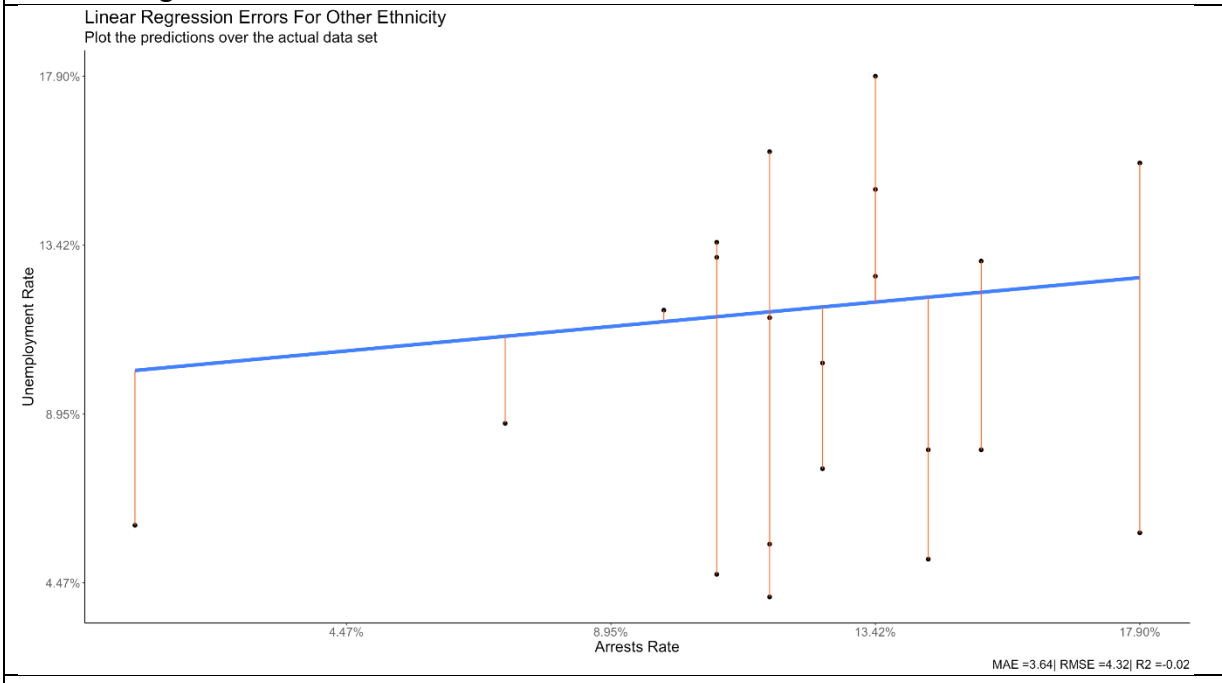
4.2.5 Other Ethnicity

The graph below plots unemployment rate against the number of arrests for Other ethnicity, by looking at the points you can see no correlation between the two- as unemployment rate increases, the number of arrests do not necessarily increase. Due to the variation in population density in different regions and that fact that any type of ethnicity, apart from those included in this project, could have been inserted into this category, there is no clear correlation.

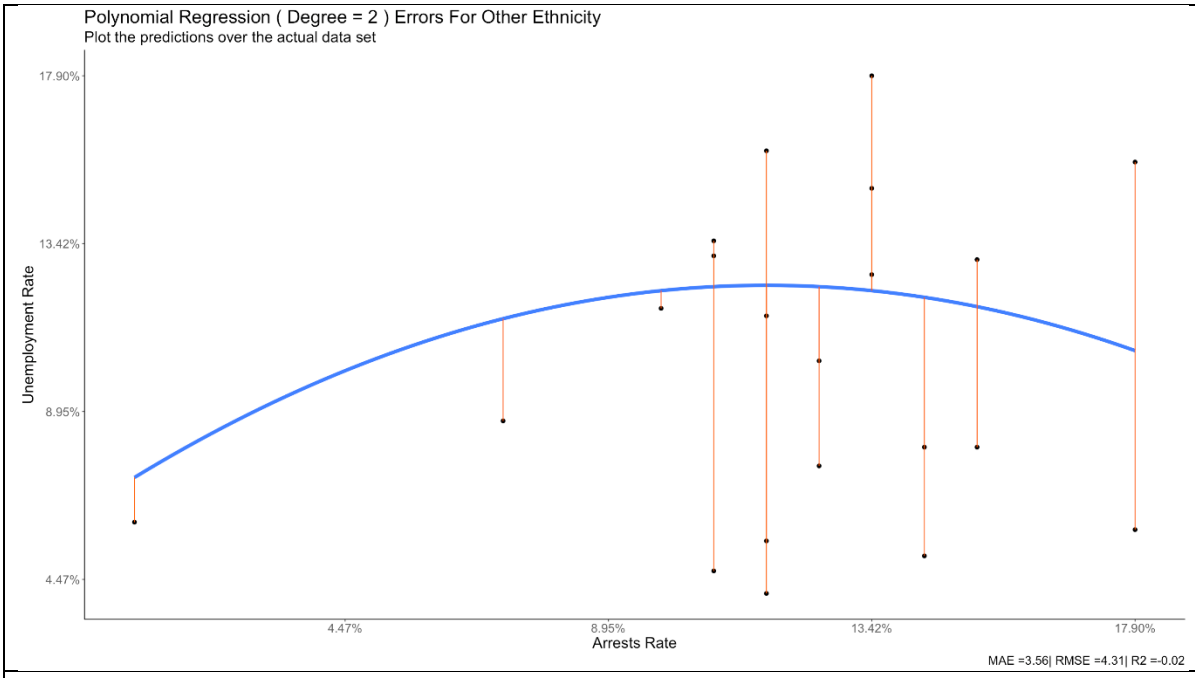
Correlation



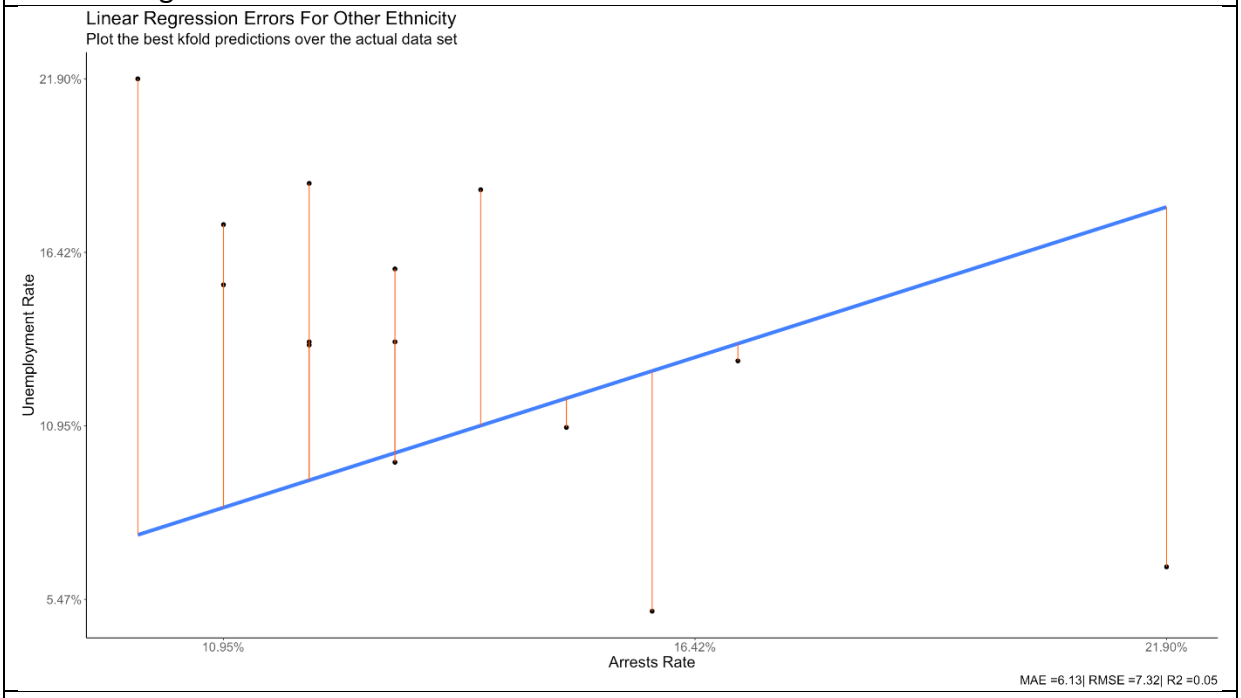
Linear using holdout



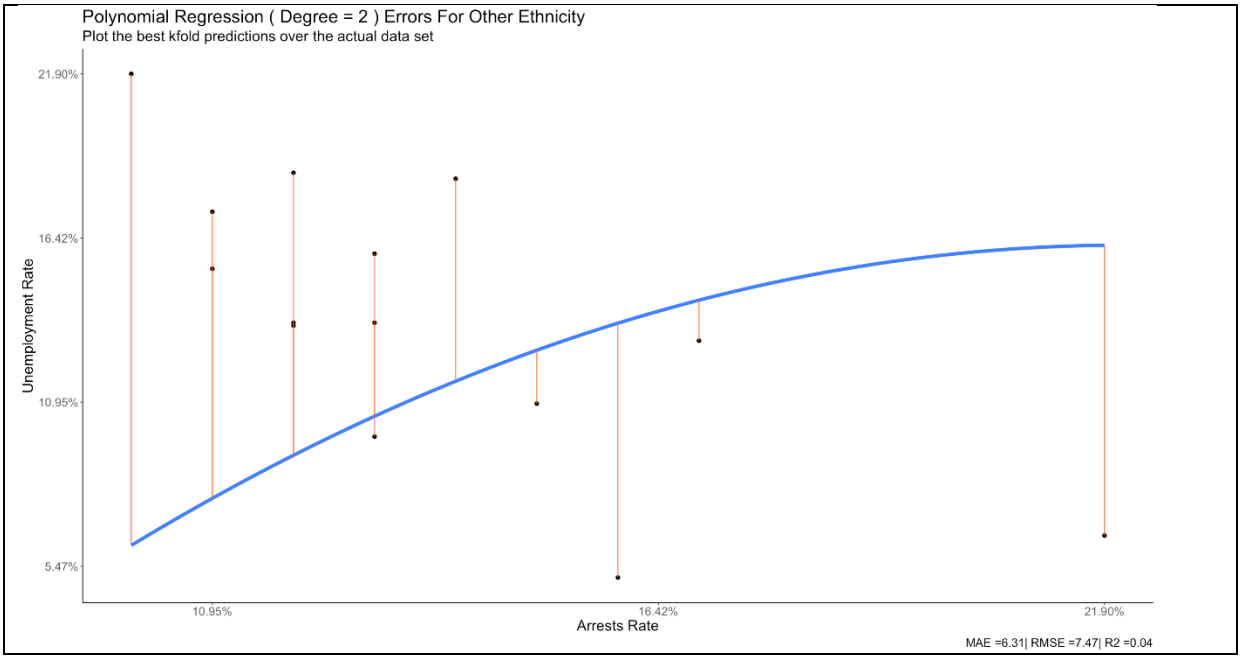
Polynomial using Holdout



Linear using KFold



Polynomial using KFold


Polynomial Regression (Degree = 2) Errors For Other Ethnicity
Plot the best kfold predictions over the actual data set
21.90%
16.42%
10.95%
5.47%
Unemployment Rate
10.95%
16.42%
21.90%
Arrests Rate
MAE =6.31| RMSE =7.47| R2 =0.04

5. Evaluation

After reviewing the field types in the datasets and finding the values were continuous, we decided that the models available to us would be linear and polynomial regression models. We knew that polynomial regression models may offer a better solution as linear regression models assume a linear correlation, however we were also aware that due to the small sample size there was an increased chance that the polynomial model would overfit the data.

Initially we used one model for all the ethnicities, however due to the difference in population density of different ethnicities we decided to split the model into 5 separate models for each ethnicity. This also made more sense for our problem, as we could get more insightful data corresponding to each ethnicity. We also initially used number of arrests, however due to the large difference in population size between ethnicities we converted this to rate of arrests (number of arrests for ethnicity / population for ethnicity) in order to get a more comprehensive model.

Due to the small sample size and consequent underrepresentation in the dataset, we decided to use K-Fold cross validation instead of using the holdout method by simply splitting the data 70/30- this way all the data would be represented. K-Fold runs through the dataset K times, splitting it into validation and train data, so all data is used in the training stage of the model, instead of plain splitting where a section of the data used to validate the model would be completely hidden from it at its training stage.

In order to test our accuracy, we used R^2, MAE and RMSE- a combination of these allows us to evaluate the model more accurately, using only one metric it becomes easy to overlook weaknesses or flaws in the model.

Algorithm	**Explanation**
$R^2 = 1 - \frac{RSS}{TSS}, where:$	$RSS = Sum\ of\ Squares\ of\ Residuals$ $TSS = Total\ Sum\ of\ Squares$
$MAE = \frac{\sum_{i=1}^{n} \lvert y_i - x_i \rvert}{n}, where:$	$y_i = prediction$ $x_i = true\ value$ $n = total\ number\ of\ data\ points$
$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)}{n}}, where:$	$x_i = true\ value$ $\hat{x}_i = prediction$ $n = total\ number\ of\ data\ points$

5.1 Result Collection and Interpretation:

Ethnicity	Measure	Holdout		K-Fold	
		Linear	Polynomial	Linear	Polynomial
Asian	R2	0.18	0.17	0.14	0.13
	MAE	2.78	2.79	2.58	2.66
	RMSE	3.43	3.43	3.23	3.37
Black	R2	0.00	-0.02	0.02	0.03
	MAE	4.19	4.07	5.23	4.55
	RMSE	5.01	4.87	6.31	5.4
Mixed	R2	0.12	0.16	0.12	0.11
	MAE	4.08	4.68	3.90	4.29
	RMSE	4.85	5.78	4.62	4.95
White	R2	0.75	0.76	0.75	0.78
	MAE	1.03	1.02	1.66	1.63
	RMSE	1.37	1.31	2.05	2.03
Other	R2	-0.02	-0.02	0.05	0.04
	MAE	3.64	3.56	6.13	6.31
	RMSE	4.32	4.31	7.32	7.47

R2 measures the extent to which one variable explains the variance of the second.

- An R2 of 1.0 is the best value. It means we have no error in our regression.
- An R2 of 0 means our regression is no better than taking the mean value, i.e., we are not using any information from the other variables.
- A Negative R2 means we are doing worse than the mean value.

K-Fold performed slightly better than holdout using this metric, on average achieving a higher R2 value, suggesting that using K-Fold creates a model that more accurately measures the relation of the two variables. When using this measure to evaluate linear against polynomial regression models we can see that the linear model tends to achieve a higher R2, and less frequently achieves a negative R2- using this as a singular metric we can ascertain that the linear model is more effective. This is due to the fact that the polynomial model tends to overfit to the data.

MAE and RMSE measure the average magnitude of the errors in a set of predictions.

- A MAE or RMSE of 0 means the model has no errors.
- MAE and RMSE are measures from 0 to infinity, the lower the value the better.
- RMSE will be larger than MAE and is useful for evaluating large errors.

Interestingly, using MAE as a metric, holdout performed more accurately than k-fold- however looking at RMSE, k-fold achieved lower measures more often, this would suggest that although k-fold produced more errors than holdout, these errors were not as severe. We did not expect our model to have high accuracy at the start, as the dataset was very

under representative, so we do not discount K-Fold as less effective from these results. On average, both the linear and polynomial achieved similar MAE and RMSE values- so it is difficult to use this metric alone to evaluate which one is more accurate.

5.2 Conclusion:

Our aim in this project was to discover whether unemployment rates affect the crime rates, within different ethnic groups. The assumption made by the group was that unemployment and arrests correlate. For this to be true, we would have needed to observe for a given ethnic group, an increase in the unemployment rate, and in the same period an increase in arrests, or vice versa.

For most of the ethnicities, this was indeed the case. Asian, Mixed and White ethnicities all displayed a positive correlation, with the latter showing the strongest relationship. This can be attributed to the fact that there is enough sample data for White ethnic people in the dataset, and so the model has sufficient data to work with. However, no real correlation was found for Black and Other ethnicities. Based on the given data, unemployment rates do not affect the arrests rates for these ethnicities. We can attribute this outcome, for Black ethnic people, to the uneven percentage of arrest for people of this group. Also, since the Other ethnicity category can include more than one ethnicity group, it could have caused the data to be vague.

Our recommendations for senior management would be that if this project were to be conducted again, linear regression is a better modelling solution to uncovering the correlations within datasets such as these. This is based on the fact that polynomial regression can over fit data of this kind, whenever there is lack of samples, which can lead to inaccurate results and thus wrong assumptions. Since this data can be considered sensitive, it is crucial to avoid this.

6. References

[1] “Number of Arrests by Ethnicity Dataset”, https://data.gov.uk/dataset/
[Online]Available:“ https://data.gov.uk/dataset/f92e60cd-ea9d-4561-b8df-ba979bda82eb/arrests-by-ethnicity ”

[2] “Unemployment by Ethnicity Dataset”, https://data.gov.uk/dataset/
[Online]Available: “https://data.gov.uk/dataset/fe6c83aa-62aa-4a8c-94cc-225f47287225/unemployment-by-ethnicity ”

[3] “Definition of Model Evaluation”, https://www.saedsayad.com/
[Online]Available:“https://www.saedsayad.com/model_evaluation.htm#:~:text=Model%20Evaluation%20is%20an%20integral,will%20work%20in%20the%20future.&text=To%20avoid%20overfitting%2C%20both%20methods,model)%20to%20evaluate%20model%20performance ”

[4] “Regression of Model Evaluation”, https://www.saedsayad.com/
[Online] Available: “https://www.saedsayad.com/model_evaluation_r.htm ”

[5] “Reference documents and Resources for Data Tables”, https://atrebas.github.io/
[Online] Available: “ https://atrebas.github.io/post/2019-03-03-datatable-dplyr/#create-example-data”

[6] “General Information about Data Tables”, https://cran.r-project.org/
[Online] Available:“https://cran.r-project.org/web/packages/data.table/vignettes/datatable-intro.html”

[7] “R Language for Data Science Sheet”, https://s3.amazonaws.com/
[Online]Available:“https://s3.amazonaws.com/assets.datacamp.com/blog_assets/datatable_Cheat_Sheet_R.pdf”

[8] “The Complete ggplot2 Tutorial”, http://r-statistics.co/
[Online]Available: “http://r-statistics.co/Complete-Ggplot2-Tutorial-Part1-With-R-Code.html"

[9] “Information for Data Transformation”, https://r4ds.had.co.nz/
[Online]Available: “https://r4ds.had.co.nz/transform.html”

[8] “Scatterplot”, https://www.r-graph-gallery.com/
[Online]Available: “https://www.r-graph-gallery.com/scatterplot.html"

[10] “Barplot”, https://www.r-graph-gallery.com/
[Online]Available: “https://www.r-graph-gallery.com/barplot.html”

[11] “Practical Business Analytics”, https://surreylearn.surrey.ac.uk/
[Online]Available: “https://surreylearn.surrey.ac.uk/d2l/home/209047”