1.Abstract:

In this task, we will design, construct, and evaluate an exploratory analysis of 3 complex datasets taken from the 2011 census in England and Wales. We will be focusing on economic activity based on gender. Along with this, we will also be performing data projections on our datasets and attempt to make our data less complex. Our main aim is to create an interactive dashboard in Tableau based on Economic activities in England and Wales based on gender. This dashboard will consist of different visualisations, making the user gain some crucial insights relevant to our data. We will use the 2011 census data available from nomisweb.co.uk. Since our data is complex, we will use some data preparation techniques in Tableau to better understand our data and perform visualisation with ease. Finally, we will also be trying two data projection algorithms on our dataset(s) and visualising the results.

2.Introduction:

Our main target for this task is to focus on three essential aspects of economic activity based on gender: Industries, Occupation, and Hours worked. We will be collecting district-wide information found on these aspects for England and Wales. This data will give us all the information on these three aspects of each district based in England and Wales based on gender. The industries dataset will consist of different industries and the number of males/females working in that industry. The occupation dataset will consist of occupation and a count of the number of males/females in that occupation in each district. Finally, the number of hours dataset will consist of the number of hours worked based on gender in each district.
One interesting thing to notice in the hours worked was that the number of hours was further divided into specific range(s) and they were labelled as part/full time based on the range of the number of hours worked. Our focus will be to prepare our vast and complex data to be manipulated easily to make visualisations for our target audience to understand and aim that the user gains valuable insights from the visualisations provided.
Furthermore, since our data will be vast and complex, so we will also be applying two different data projections algorithms: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding(t-SNE) datasets. These algorithms reduce the complexity by experimenting with the dimensions of the data, and finally, we will be showing visualisations of our newly projected data.

3.Data Preparation and Abstraction:

On downloading our data, our dataset had a vast number of columns. There was a particular pattern that was noticed. For example, Our Industries dataset starts with a total count of the number of people column first, followed by the count of all genders for each industry. Lastly, count the number of people working in an industry based on gender. Each row consisted of one district. The same pattern was noticed in the other two datasets as well.

	B	C	D	E	F	G	H	
	geography	geography	Rural Urba	Sex: All pe	Sex: All pe	Sex: All pe	Sex: All pe	Sex
1	Darlingtor	E06000005	Total	49014	303	117	4900	
1	County Du	E06000047	Total	227894	2124	922	29983	
1	Hartlepoo	E06000001	Total	37767	139	267	4410	
1	Middlesbr	E06000002	Total	54547	82	387	4430	
1	Northumb	E06000057	Total	146901	3739	770	13569	
1	Redcar an	E06000003	Total	56354	285	1126	5815	
1	Stockton-	E06000004	Total	87122	221	617	8672	
1	Gateshea	E08000037	Total	91877	153	160	9547	
1	Newcastle	E08000021	Total	119335	153	231	7315	
1	North Tyn	E08000022	Total	96026	127	393	7883	

One more challenge faced was that the column name had ‘separators’ like ‘;’, ‘:’,’,’,’.’. This is certainly a problem as it can provide us difficulty to a certain degree while performing operations for creating visualisations on the Tableau.

Sex: Males; Industry: M Professional, scientific and technical activities; measures: Value

AJ	AK	AL	AM	AN	AO	AP	AQ	AR
Sex: Male	Sex: Male	Sex: Male	Sex: Male	Sex: Male	Sex: Male	Sex: Male	Sex: Fema	Sex: Fema
202	1284	1119	2019	1212	1581	752	23750	82
1251	4803	5131	9721	7452	6278	3898	109422	502
268	951	830	1238	929	1030	619	18256	25
316	1406	1409	1537	1485	1879	968	26350	30
1091	4236	3239	7099	3656	4921	2748	71053	853
417	1586	1320	1715	1415	1378	917	27635	63

Due to our data being in such a format and gender being our focus, our data turns out to be huge and has a lot of columns. Due to such complexity, we cannot create the visualisation that we intend to.
Hence, we will prepare and manipulate our data to reduce the number of columns consisting of our economic factor. Occupation/Industry/Hours worked our social factor that is gender and the count. Our rows should contain the district name, value for economic factor, gender (male/female), and that specific district’s total count.

District name	Economic activity	Gender	count
A	B	Male	XYZ
A	B	Female	XYZ
A	B	Male	XYZ

Firstly, we can reduce the complexity of data by removing the total count and columns consisting of both genders’ total values. We will reduce the number of columns in our data by doing this. We should not concern ourselves with whether any data was lost as the sum of the rest of the columns divided based on gender is equal to the total count of columns.
Next, we will be using the “pivot” and “split” functions in Tableau for data preparation. Data pivoting enables us to rearrange the columns and rows in a report so we can view data from different perspectives. We can pivot all the Economic activities rows into one column and their values into another column using the pivot function.
However, our job did not get done there. Each column was named with specific separators, which gives us more information. While pivoting the data, the whole column is printed as a row. This means that the separators are still there. We need to split our data even further to gain more helpful information. We cannot just rely on Tableau’s splits as our data consists of many separators as delimiters. Hence, we will be using the custom split option provided by Tableau to split data accordingly. Once the data is split, we will remove the unnecessary columns and keep the relevant columns. We will then further format our columns accordingly.

Hours worked after data prep Geography	Hours worked after data prep Geography Code	Hours worked after data prep Part/Full- time	Hours worked after data prep Sex	Hours worked after data prep Time(hrs)	Hours worked after data prep Count
Darlington	E06000005	Full-time	Females	31 to 48 hours worked	11,817
Darlington	E06000005	Full-time	Females	49 or more hours worked	1,203
Darlington	E06000005	Part-time	Females	15 hours or less worked	2,834
Darlington	E06000005	Part-time	Females	16 to 30 hours worked	7,896
Darlington	E06000005	Full-time	Males	31 to 48 hours worked	17,463
Darlington	E06000005	Full-time	Males	49 or more hours worked	4,198
Darlington	E06000005	Part-time	Males	15 hours or less worked	1,259
Darlington	E06000005	Part-time	Males	16 to 30 hours worked	2,344
County Durham	E06000047	Full-time	Females	31 to 48 hours worked	56,431

Fig: How our data looks after performing the pivot function and splitting the data accordingly

4.Task Definition:

Before we begin our visualisations, we need to ask critical questions, 1. *Why* is this task performed? 2. *How* should this task be performed? 3. *What* comprises a visualisation, or what does it pertain? We will be answering these questions in this section. The visualisation tasks are performed to make the user gain insights into the information presented to them through the visualisations provided. We aim to present an enjoyable visualisation to our target audience: people with moderate to no information about the topics we are presenting. We also want our users to discover new information by looking at our visualisations. Clearing our data gives us insights into our data. We will be focusing on answering specific questions for each dataset as follows:

1. **Occupation:**
   - Which occupation(s) are dominated gender-wise?
   - How do different occupation(s) occur based on districts, and what could be the reason behind this?
2. **Industry:**
   - Which industries are dominated based on gender?
   - Which industries do people work more in are based on the location of districts?
   - Which industries are more in which district, and what could be the reason behind that?
3. **Hours Worked:**
   - Which gender works more part/full time?
   - What is defined as ‘part/full time’ based on hours worked.
   - Which has more frequency between part-time and full time.

Coming to the how part, we will be using different visual and encoding and interaction techniques with the help of Tableau. We will also be using additional filters and colouring that are pleasing to the human eye so that the user can gain information on some regions of their choice. For example, if the user wants to focus on the occupations dominated by females district wise, he can do that using these filters. This filtering can be a reason for what this task will pertain. The main area to focus on the “what” part is that we want our visualisation so that if a user had a specific idea/input for a particular domain, they could gain insights on that using these filters.

Our visualisation aims to show relation/insights between gender against economic activities like the occupation of the people, the industries people work in, and the number of hours people works in districts wise for England and Wales. So, our idea is to create a dashboard of the Whole UK map

portioned district-wise and the plots of the abovementioned relations. The UK map will be used as a filter. If the user wants to gain information about a particular district, they can click the district. Our visualisations will change, showing the data according to the location selected.

5.Visualisation Justification:

On our first sheet, we draw out the map of the UK, which includes England and Wales based on districts. We then colour the map out using the count of people working in industries. This was to give the user an idea of which districts are more populated and which districts aren't. The colour scheme I used was orange. The reason to go with this colour was that most of my plots and dashboard followed the theme of orange and white. In this map, the district and intensity of the orange colour were proportional. Suppose the district had more population, which implied that the colour on the map for that district was darker.

Similarly, the district with less population had a lighter shade of orange on them. A filter option was provided so that the user could understand what different shades of orange meant. In addition to that, we also filtered the count of people so that the users can interact and check the districts and their count for a range of their own choice.

For our second sheet, we focused on the number of hours an individual work. I have decided to plot the count of people (our continuous values) against the columns part/full time, time (hrs) and sex. I chose to use a stacked bard in a horizontal representation. The reason for doing this was if you look at the plot, it gives precise information at different levels. First, it focuses on whether people are working part-time or full time. Then it focuses on the range of the number of hours people work, and finally, it focuses on gender. The gender bars are colour coded. This gives the user clarity over males and females. I have also labelled which colour represents which bar. I have also filtered the rows we consider, i.e., part-time/full time, sex. The count of people has also been filtered. This allows the user to interact with our plots and filter data of their own choice. For example, if the user wants to understand the count of female part-time workers, they can achieve that using this filter.

For my next sheet (3), I have decided to plot a treemap of the occupation of people based on sex. The reason to use it was because each industry was a category, and this looked like a proper structure to represent all the industries. In addition to that, I have also used a text mark to display the count under each industry. The occupation count was also coloured in orange, as this is the theme we are going with. The darker the colour, the more the count of people. The reason to use this was to make the user understand which industry has more count of which gender. This will make the user gain information about the sectors dominated by females and the sectors dominated by males. For example, we can tell by looking at the graph that females were dominating the administrative and secretarial occupations.

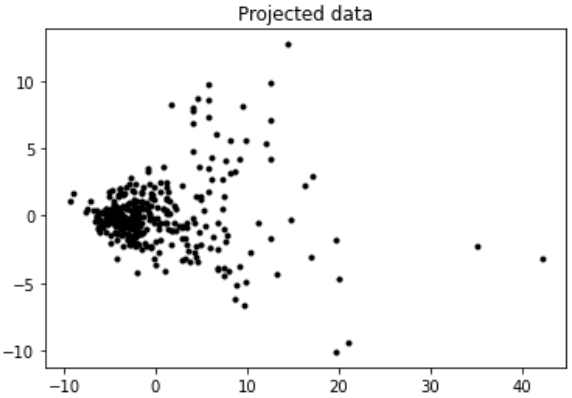
In contrast, males dominated the skilled trades sector and some occupations like professional jobs, the count was in a similar range, and there was no bias. In addition to this, I have also filtered the occupation and sex columns. The reason to do this was so that the user could interact with the plots and gain insights into their choice. For example, if the user wants to know the count of the female professional in a particular district, they can get that information using the filters provided.

In our final sheet (4), I have decided to have a bar plot of gender and count of people against the industries. The reason to do this was that it gives a nice comparison plot of industries based on gender. The user can see the plots and gain insights based on gender. The bars can make the user

distinguish which industries have more count of males/females. I have also coloured the plots orange to follow the theme. A text displays the number of people working in the industries at the top of the plot so users can see the count of people working in the industries instead of hovering over it. The gender aspect and the industries are filtered. The reason for doing this was to make the users interact with the plots. For example, if the user wants to find the count of females in the education industry for a particular district, they can do that using these filters.
Finally, all these sheets are made into a dashboard where the map of the UK is our main sheet in the centre and is used as a filter. This enables the user to find information district wise. I have placed the occupation treemap on the left-hand side, and on the right-hand side of the map, I have set the industry and hours worked plot. The reason to do this was that the treemap was enormous compared to others. All the filters mentioned in the sheets have also been placed next to the visualisations so that the user can interact as much as they like to gain new and different types of information.

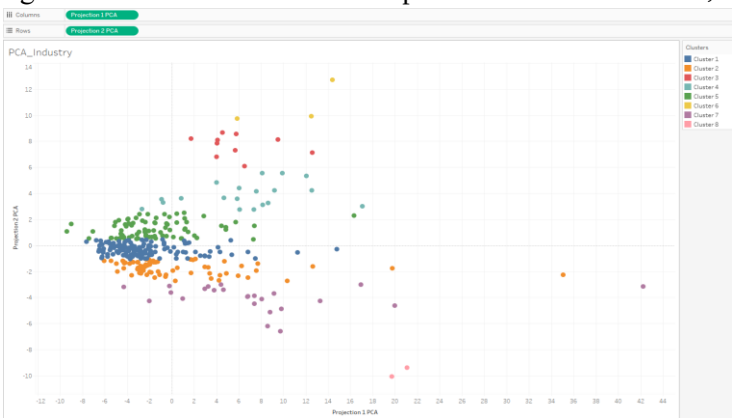
Note: - This dashboard may look different based on the user display size and resolution.

Coming to our data projection part, our dataset contains a lot of datasets. A model with too many degrees of freedom is likely to overfit the training dataset and may not perform well on new data. The point of data projections/ dimensional reduction is to reduce the dimensions of our data so that it has fewer input variables and generalises to unseen data. We usually do this in the data pre-processing part of machine learning. I decided to use two different data projection algorithms on two other datasets and then plotted our results in Tableau. My first pre-processing step for both datasets was to drop the data with a categorical value, like the area name. I noticed that our data wasn't scaled correctly, so I began scaling my target columns using the StandardScaler from the Sklearn library. Once the data was scaled and ready to use, I then applied my data projection algorithms to them. The first algorithm I used was Principal Component Analysis (PCA) for my industry dataset. This data set contained 60 columns, and obviously, these columns were too high. PCA is a method for lowering the number of dimensions in a dataset while keeping most of the data. I first computed the eigenvalue and eigenvectors. I used these eigenvalues and eigenvectors to compute a threshold plot to determine the optimal number of components having a threshold variance above 95%. From this plot, nine components met the criteria of threshold variance. We applied PCA on our scaled data with my "n_components" setting equal to 9. After fitting the data, we store our data projections' first and second results and plot those results to understand our new dimensions.



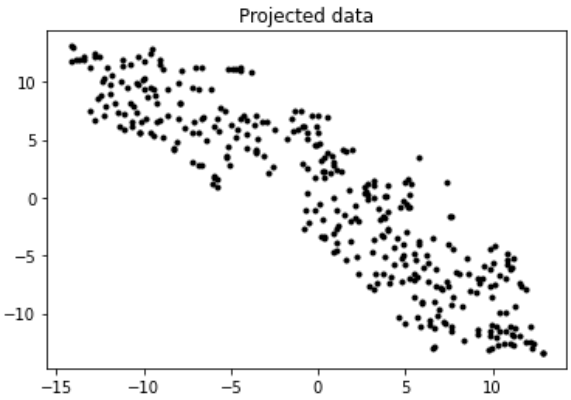
But our results are in a single colour. Since labels are absent, we couldn't tell the count of our new reduced dimensions. To understand the unique dimensions formed, we export the first two results of our fitted model into a data frame and export it to Tableau. Only two results/components were considered because we cannot

plot greater than two dimensions in Tableau. We use Tableau as Tableau is advanced enough to segregate the features into clusters using the "cluster" feature. Then we plot the results in Tableau, which are as follows:

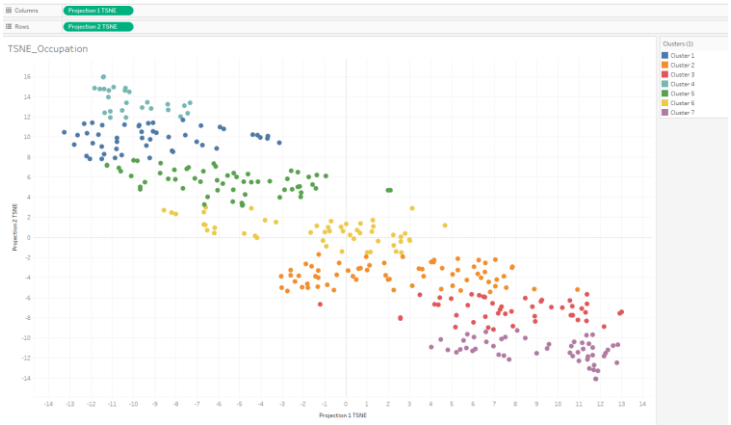


The cluster on our plot tells us that the 8 clusters are our new reduced dimensions, originally 60. More precisely, we can group the local authority regions into eight groups/clusters given the industry by gender. One more interesting thing to notice is that some of our clusters have outliers.

The following algorithm I used was "T-Distributed Stochastic Neighbouring Entities (t-SNE)" on our occupation dataset. This dataset contains 33 columns. Another technique for dimensionality reduction is t-Distributed Stochastic Neighbour Embedding (t-SNE), which is particularly well suited for displaying high-dimensional datasets. Unlike PCA, it is a probabilistic technique rather than a mathematical one. The only disadvantage of t-SNE is Since t-SNE scales quadratically in the number of objects N. Its applicability is limited to data sets with only a few thousand input objects; beyond that, learning becomes too slow to be practical (and the memory requirements become too large). The method was the same as PCA, where we first scaled our data and then applied the algorithm to our new scaled algorithm. Then we plotted the first two results of our model to understand our new dimensions



But our results are in a single colour. Since labels are absent, we couldn't tell the count of our new reduced dimensions. To understand the unique dimensions formed, we export the first two results of our fitted model into a data frame and export it to Tableau. We use Tableau as Tableau is advanced enough to segregate the features into clusters using the "cluster" option. Then we plot the results in Tableau, which are as follows:



The cluster on our plots tells us that the 7 clusters are our new reduced dimensions, originally 33. To be more precise, we can group the local authority regions into seven groups/clusters given the occupation by gender

6.Conclusion:

From performing all the visualisation above, we can conclude that different genders dominate different occupations and industries. We can note that men were involved in more physical and labour/skilled based work like construction, mining etc. In contrast, females dominated industries like education, health, social services, etc. We can also notice that people worked full-time more than part-time. The count of females was more than males when it came to full-time work. If we had to guess, the proportion of part-time was less mainly because people in the community, like students, took part-time jobs. One more good insight we got was the district close to water bodies had an agriculture industry dominated. Major cities of the UK had more population and industries than the outskirts and suburban areas.

From our data projections, we can understand that sometimes having these many columns/dimensions in our data can increase the complexity of our data which can cause our machine learning models to overfitting and perform poorly on unseen data. By applying data projections, we could notice that the dimensions of our data are reduced, which also reduces the complexity of our data. This makes our task of performing machine learning algorithms easier as there are fewer input variables, making our model a good fit for generalising.

Thank You

References:

- https://towardsdatascience.com/visualising-high-dimensional-datasets-using-pca-and-t-sne-in-python-8ef87e7915b
- https://machinelearningmastery.com/dimensionality-reduction-algorithms-with-python/