Salman Bhutta - 13030025
Hassan Tariq – 13030023

Data Understanding Report

Phase 2

Table of Contents

1. Overview:

As our selected Dataset help us to predict salary for jobs posted in United Kingdom, It contains data collected from different companies from different states of UK, salary of half of the jobs are publicly displayed in advertisement across UK. So our prediction will enhance job search for both job seekers and companies, It will help job seekers to look out for jobs relevant to their desired salary and will help employers to find right person for different positions.

The data set contains 244,768 records, which are mostly unstructured text, with a few structured data fields, as these records were collected from different sources from different regions. Successful models will incorporate some analysis of the impact of including different keywords or phrases, as well as making use of the structured data fields like location, hours or company. Some of the structured data shown (such as category) is 'inferred, based on where an advertisement came from or its contents, and may not be "correct" but is representative of the real data.

Attribute Name	Column Description
Id	Unique record identifier
Title	Column contains information about job title
FullDescription	Column contains information about job description
LocationRaw	Column contains information about job location
LocationNormalized	Column contains normalized information about job location
ContractType	Column contains information about job specific contract
ContractTime	Column contains information about time allocated for specific job
Company	Column contains information about job posting company
Category	Column contains information about job type
SalaryRaw	Column contains information about job salary range
SalaryNormalized	Column contains normalized information about job salary
SourceName	Column contains data source name

As we have more than 200k records of normalized salaries posted against various job, by using our prediction model, job seeker will get most relevant available jobs according to different filters provided by input user. Different filters will include interesting candidate job region, nature of job, time and type of contract, along with title of job. We will design an interactive front end interface for job seeker to look out for his/her desired jobs according to various provided filters.

2. Data Pre Processing:

Data Attributes along with their data types and missing values count given in the table which will help decide the columns to use in our prediction model.

Attribute Name	Attribute Type	Missing Count
Id	integer	0
Title	polynominal	1
FullDescription	polynominal	0
LocationRaw	polynominal	0
LocationNormalized	polynominal	0
ContractType	binominal	179,326
ContractTime	binominal	63,905
Company	polynominal	32,430
Category	polynominal	0
SalaryRaw	polynominal	0
SalaryNormalized	integer	0
SourceName	polynominal	1

a. Feature Selection:

Attribute Name	Missing Values	IS Part of Prediction Model	Reason
Id	NO	NO	ID is Associated with every Title, we will use Title instead as our prediction criterion as the user is aware of the Job title and not the ID.
Title	NO	YES	Title will play a key role in our salary prediction model as most of the users will try finding salary using the Job title.
FullDescription	NO	NO	
LocationRaw	NO	NO	We will need location in our prediction model, but we will be using normalized location
LocationNormalized	NO	YES	Normalized location will help predict salary according to the location / state stated by user according to the user requirement
ContractType	YES	NO	Contract type could have been helpful in our prediction model but the count of missing values forced us to exclude this attribute. Missing values were almost 75%

			and averaging the attribute to fill the missing count would result in biased outcome.
ContractTime	YES	YES	Contract time which is a binary attribute will help user choose the type of job desired and will predict the salary accordingly. Here we will also analyze which job type permanent or contractual is better paid in the United Kingdom.
Company	YES	YES	We would use Company as a feature for our prediction model as this will assist the job seeker to view salaries of desired job in a specific company.
Category	NO	YES	Along with title, company will be the core feature of our prediction model. Statistics show that data contains 29 different job categories. This will help easily identify the area of search along with the title.
SalaryRaw	NO	NO	As an input to the prediction model, this feature will lead us to our goal of optimal solution. Since we have normalized salary available, we will use that in our model.
SalaryNormalized	NO	YES	As an input to the prediction model, this feature will lead us to our goal of optimal solution. Given different input attributes, salary will be our output based on the provided salary to the model.
SourceName	NO	NO	Provided data has been collected from multiple sources across United Kingdom. We need this to judge authenticity of data but this is not required in our prediction model.

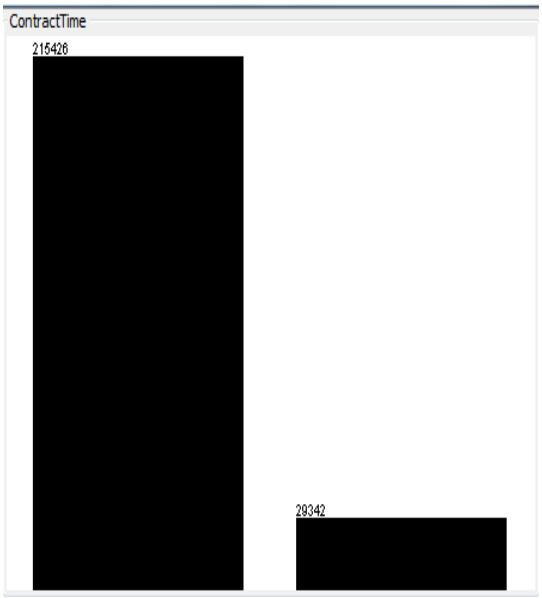
b. Removing Missing Values:

Attribute Name	Missing Count	Action Taken
ContractType	179,326	Missing Value count approx. 75% therefore column has been discarded for use in our prediction model
ContractTime	63,905	Missing values have been filled by taking the average of the values available for this attribute
Company	32,430	Missing values have been filled by taking the average of the values available for this attribute

3. Understanding the Data:

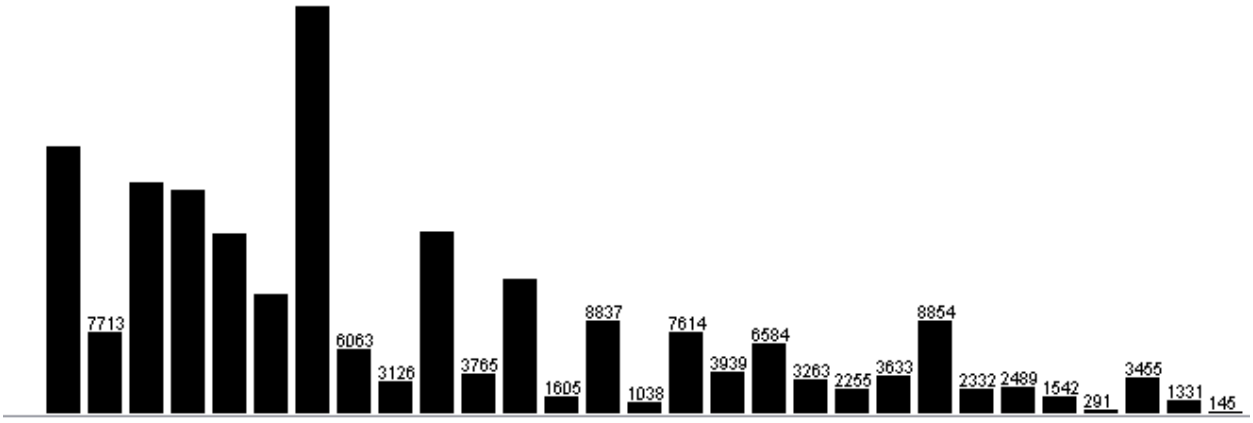
Attribute	Distinct Values
Title	135,431
Location	2,732
Contract Time	2
Company	20,812
Category	29

a. Contract Time:



Permanent	Contractual
215,426	29,342

b. Category:

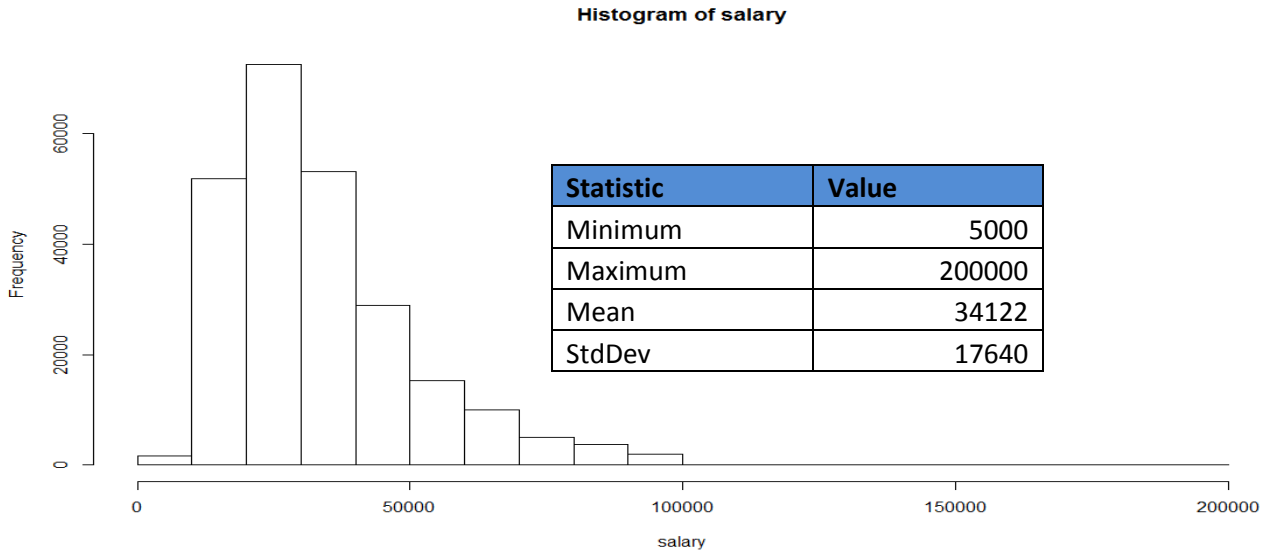


c. Categories Normalized:

No.	Label	Count
1	Engineering Jobs	25174
2	HR & Recruitment Jobs	7713
3	Accounting & Finance Jobs	21846
4	Healthcare & Nursing Jobs	21076
5	Other/General Jobs	17055
6	Hospitality & Catering Jobs	11351
7	IT Jobs	38483
8	Customer Services Jobs	6063
9	Travel Jobs	3126
10	Sales Jobs	17272
11	Manufacturing Jobs	3765
12	Teaching Jobs	12637
13	Creative & Design Jobs	1605
14	Trade & Construction Jobs	8837
15	Property Jobs	1038
16	Admin Jobs	7614
17	Legal Jobs	3939
18	Retail Jobs	6584
19	Consultancy Jobs	3263
20	Energy, Oil & Gas Jobs	2255
21	Logistics & Warehouse Jobs	3633
22	PR, Advertising & Marketing Jobs	8854
23	Charity & Voluntary Jobs	2332
24	Scientific & QA Jobs	2489
25	Maintenance Jobs	1542
26	Domestic help & Cleaning Jobs	291
27	Social work Jobs	3455
28	Graduate Jobs	1331
29	Part time Jobs	145

From large amount of data, we have been able to identify these 29 Job categories. All Job titles belong to one of these Categories. This will act as a major input criterion to narrow our search provided 135,431 distinct Job Titles.

Salary Normalized: Peak shows that mostly job salaries lie between 30-35k marks annually.



Statistic	Value
Minimum	5000
Maximum	200000
Mean	34122
StdDev	17640

Salary Variation with Outlier Analysis