Uncleaned Data Science Dataset

Issues with Uncleaned Data Science Dataset

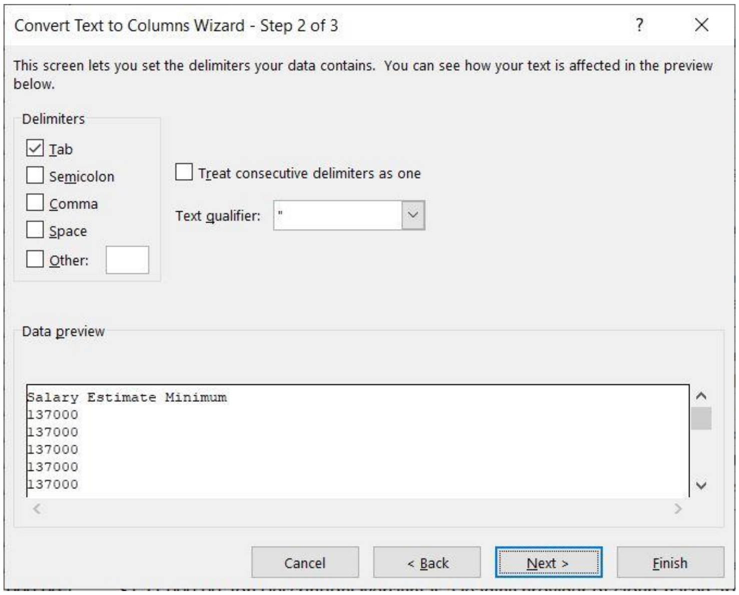
1. Incorrect Datatypes.
   The Column [Salary Estimate] has a data type of text which could make complexity in the further steps like analysis. It should be converted to a currency data type for data visualization.

2. Missing/Unknown Values
   There are unknown or null values in the columns [Sector], [Industry], [Competition] and [Year]. These unknown values cannot yield a quality analysis. So, we need to clear these unknown value records if it does not affect the data.

Techniques used to overcome the issues

I have used Microsoft Excel and Power Query to overcome above mentioned issues.

1. **Solving incorrect data type problems:**
   To overcome incorrect datatype values which is in text for currency values we need to change the data type value to currency. In my case the data was combination of text and number, there was a range $137K-$171K, So I used delimiter tool to split it into 2 columns. The result happened to be $137K and $171K in different columns, then I replaced $ with none and K with 000 respectively, to get my desired value of 137,000 and 171,000. Ultimately converting the 2 columns data types to currency.



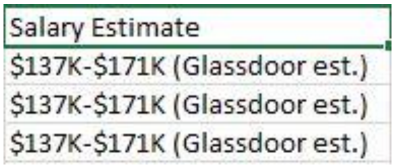
This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

☑ Tab
☐ Semicolon
☐ Comma
☐ Space
☐ Other:

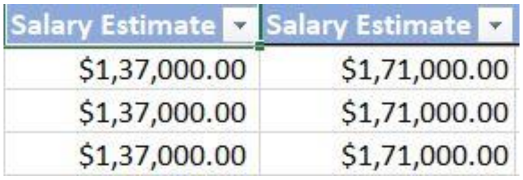
☐ Treat consecutive delimiters as one

Text qualifier: "

```
Salary Estimate Minimum
137000
137000
137000
137000
137000
```



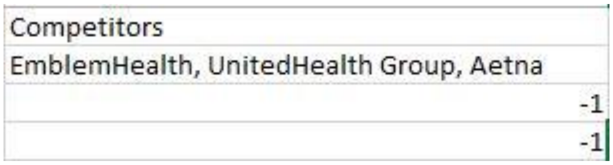
Salary Estimate
$137K-$171K (Glassdoor est.)
$137K-$171K (Glassdoor est.)
$137K-$171K (Glassdoor est.)



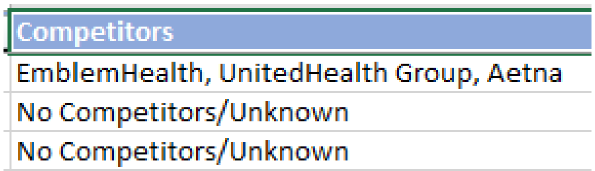
Salary Estimate	Salary Estimate
$1,37,000.00	$1,71,000.00
$1,37,000.00	$1,71,000.00
$1,37,000.00	$1,71,000.00

2. **Solving missing/unknown values:**
   There was a less than a 100 unknown values (-1) in the columns [Sector] and [Industry]. Therefore, we could remove these unknown value records since it does not affect the data which consists of more than 600 rows.

   I have used Find & replace tool to clear the unknown values and replace it with meaningful data.



Competitors
EmblemHealth, UnitedHealth Group, Aetna
-1
-1



Competitors
EmblemHealth, UnitedHealth Group, Aetna
No Competitors/Unknown
No Competitors/Unknown

Uncleaned data:

15 columns and 672 rows

Cleaned data:

15 columns and 594 rows

Software used: Microsoft Excel

Assignment by:

Varun Kumar Kolloju

Matriculation number 95675739