Assessment 1 Template: Dataset preparation report

Introduction

This report documents the data preparation process using data gathered from various HR systems used by Revolution Consulting. Data preparation plays a key role in the data science pipeline as it identifies and removes errors that may cause anomalies in the later stages of data exploration, data modelling and data presentation. Revolution Consulting, an IT Consulting firm, has been facing issues with declining work quality produced by its consultants. The company is facing higher turnover in consultants and newer consultants are not as knowledgeable or skilled. This report documents the preparation process and provides initial data exploration to assist with further analysis in determining the underlying issues causing staff at Revolution Consulting to leave.

Data preparation

Overview

The data is a structured data set of employee information compiled from various HR systems and surveys used by Revolution Consulting. Consists of numerous data features that are nominal, ordinal and interval. The relationships between the data features can be analysed in identifying the underlying causes for the IT firm's staff retention and work quality issues.

Process

1. *Using a numbered list, provide the steps you used to prepare your data*
2. *Ensure you give a justification for any choices made at each step*
3. *Review the case study and the Assessment brief (including the rubric) for more guidance*

Issues discovered

Use the table below to record your issues.

#	Issue name	Location	Code to identify	Rationale and solution
1	*Correct data types for columns*	*Age, BusinessTravel, Business Unit, Gender, Marital Status, Overtime, Resigned*	*revolution_df is name assigned to data frame:* *revolution_df.info()*	*"Age" set to integer as it is numerical without decimals, Overtime and Resigned changed to "1" and "0" with integer data types as they have binary responses useful for data modelling. Other columns changed to 'Category' as suitable for 'objects'*
2	Convert non-numeric columns to upper case.	Resigned, Business Travel, Business Unit, OverTime, Gender, Marital Status columns	revolution_df['Resigned'] revolution_df['BusinessTravel'] revolution_df['BusinessUnit'] revolution_df['OverTime'] revolution_df['Gender'] revolution_df['MaritalStatus']	revolution_df[column].str.upper() makes responses in non-numeric uniform and assists with cleaning

3	Age – data entered as 36a	Age column of data frame	revolution_df['Age'].value_counts()	36a is a typo Solution: revolution_df['Age'].replace('36a', '36', inplace = True)
4	Different inputs in resigned column, white spaces	Resigned Column	revolution_df['Resigned'].value_counts()	Stripped responses of white spaces and made all responses upper case and uniform
5	Different inputs in Business Travel, white spaces	Business travel column	revolution_df['BusinessTravel'].value_counts()	Stripped responses of white spaces, all responses made uppercase and uniform using .replace function
6	Entry in Business Unit entered incorrectly as ‘female’	Business Unit Column	revolution_df['BusinessUnit'].value_counts()	Filtered row using: revolution_df[['BusinessUnit']][revolution_df.BusinessUnit == 'FEMALE'] Used replace function to correct: replace('FEMALE' , 'SALES', inplace = True)
7	Incorrect entry in Gender	Gender Column	revolution_df['Gender'].value_counts()	Filtered row using: revolution_df[['Gender']][revolution_df.Gender == 'SALES'] used replace function to correct: revolution_df['Gender'].replace('SALES', 'FEMALE', inplace = True)
8	Cleaning responses in Gender column,	Gender Column	revolution_df['Gender'].value_counts()	Remove white spaces with strip function. Used replace function to make responses uniform
9	Missing Data	Education Level,	revolution_df['EducationLevel'] = revolution_df['EducationLevel'].fillna(-1)	Calculated median of 3 and replaced for row: revolution_df.loc[1215, 'EducationLevel'] = 3
10	Inconsistent responses in Marital status	Marital Status column	print(revolution_df['MaritalStatus'].value_counts())	Removed white spaces using strip function. Used replace function to make responses uniform.
11	Missing info in Monthly Income	Monthly Income column	revolution_df['MonthlyIncome'] = revolution_df['MonthlyIncome'].fillna(-1)	Returned values of rows with missing columns, calculated mean income for rows according to employee’s business unit.

			no_money = (revolution_df[revolution_df['MonthlyIncome'] == -1])	Used replace function to fill missing data.
12	Overtime data cleaning	Overtime column	revolution_df['OverTime'].value_counts()	Removed white spaces. Created mask to search for any incorrect entries between overtime noted and Weekly Hours. Used replace function to correct entries
13	Overtime Missing data	OverTime Column	revolution_df['OverTime'] = revolution_df['OverTime'].fillna(-1) print(revolution_df[revolution_df['OverTime'] == -1])	Using returned employee ID rows, determined overtime based on work over 40 hours per week. If > 40, overtime replaced with yes. If hours < 40, overtime replaced with No.
14	Average Weekly Hours Worked – incorrectly entered as 400 hours	Average Weekly Hours Worked Column	revolution_df['AverageWeeklyHoursWorked'].max()	Found employee ID using mask, Checked if employee noted as doing overtime and replaced weekly hours value using median weekly hours which was 40.
15	Work Life Balance Missing	Work Life Balance Column	revolution_df['WorkLifeBalance'] = revolution_df['WorkLifeBalance'].fillna(-1) revolution_df[revolution_df['WorkLifeBalance'] == -1])	Replaced value with -1
16	Total Working years incorrect	Total Working Years	revolution_df[['TotalWorkingYears', 'YearsAtCompany']][revolution_df.TotalWorkingYears < revolution_df.YearsAtCompany]	Used loc function to locate employee, determined Total Working years to be 1, based on Years with Current Manager and Years with company to be 1.
17	Resigned Data clean	Resigned	revolution_df['Resigned'].value_counts()	Used strip function to remove white spaces. Used replace function to make responses Uniform
18	Resigned Missing Data	Resigned	revolution_df['Resigned'] = revolution_df['Resigned'].fillna(-1) revolution_df[revolution_df['Resigned'] == -1]	Dropped rows with missing values as Resigned is the Target data for our analysis.
19	Column names changed to upper case	All column titles	revolution_df.rename(columns=str.upper)	Made column names upper case as per assignment specifications.

Data exploration

Overview

Data exploration was aimed at looking to gain insights into underlying issues with staff retention and output quality facing Revolution Consulting. I chose to start by looking at proportion of business unit size to gain a better understanding of the firms operations. Next we looked at Years in Role, Job Satisfaction and Monthly Income in testing for any relationships that would provide insight into the company's staff and quality issues.

Process

Observations

#	Observations	Significance
1	**Pie Chart – Business Unit Break-up:** Shows how the company is split up amongst business units. We see that consultants make up the majority of staff, followed by Sales and lastly by Business Operations.	Though the company is facing issues with staff retention across the company, this is significant, as the company's main operations are from consulting engagements. This ties into the goal of identifying the underlying issues with consultant output quality and retention.
2	**Histograph – Job Satisfaction:** Companywide, 463 employees rate their satisfaction at a level of 4, 447 employees rate at a level of 3, 280 employees at a level of 2, and 289 employees at a level of 1.	This is an important observation, as with issues relating to staff retention and engagement, we assume that one reason an employee may leave a company would be level of job satisfaction
3	**Histograph – Years in Role:** In this graph we can observe the number of years employees have been with the company. The chart shows that 245 employees have started with in the last year (year 0). Only 59 employees have been with the company for one year. 373 employees have been with the company for 3 years. This figure reduces until year 7 where 223 people have been in the company for that time.	From the graph, we can see that many of the employees at Revolution Consulting are new. The importance of this observation is that coupled with other data features, we can get more insights into why employees would stay or leave their employer. This graph indicates that the company is having trouble keeping more experienced staff. This could relate to other pieces of data like Monthly Income, Job Satisfaction, the time spent in a role and work-life balance.
4	**Resigned by Age – Box Plot:** The box plot shows median age of staff who have resigned is 32, compared to the median of current staff which is 36. The Box plot shows the minimum age at 18 for both 'Current' and 'Resigned'. Maximum ages shown are 60 years for 'Current', and 55 years for 'Resigned'	We can see that the ages of employees resigning are lower than those staying. As such, reasons for leaving can be due to reasons that have a correlation to other features in the data set such as Job satisfaction, Monthly Income and work life balance. It also relates to the above histography, (#3) where the distribution of employees by years spend in role suggests they are having trouble retaining experienced staff. This may also be a factor in why output quality has fallen.
5	**Resigned vs Monthly Income - Bar Graph:** Shows the mean monthly income between staff who are current and who have resigned. Monthly Income (mean) for current staff is $6818.46. Monthly Income (mean) for resigned staff is $4779.16	This is a significant observation as it matches Revolution Consulting's hypothesis that they are having trouble maintaining their best staff and losing them to competing firms. It raises questions regarding how staff might view remuneration and job prospects in the firm. This has flow on effects on the declining output quality of the firms consulting team.
6	**Years in Role and Job Satisfaction – Bar Graph:** Graph compares median level of Job Satisfaction to the number of years employees have been in a role.	With the exception of dips in years 6, 12 and 13. A high rating of 5 in year 17.The graph shows that the median level of job satisfaction stays relatively consistent across the number of years in a role. We have seen in graph#3 that the majority of employees at Revolution Consulting have not been with the company for more than 4 years. This would suggest that Job Satisfaction would have very little correlation with Revolution Consulting's retention issues.

Plots

Business Unit Pie Chart: Included to visualise split of business units and understand business's operations. In this case which is consulting.

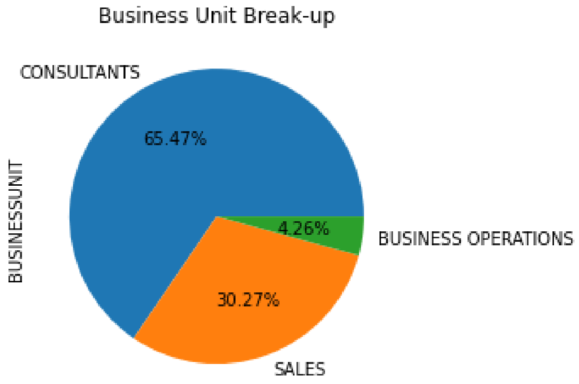
Job Satisfaction Histograph: Job satisfaction ratings provided by employees. Job satisfaction would be considered a key factor in analysing any issues with staff retention. Produced to see the distribution in job satisfaction among employees.

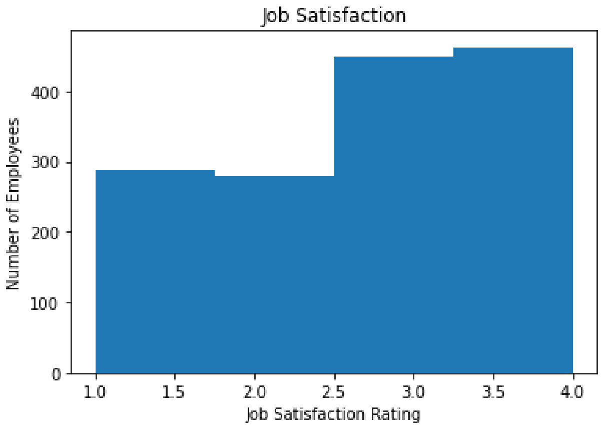
Years in Role Histograph: Depicts distribution of staff and number of years in role. Investigating if firm can retain staff for long-term, or if people are likely to leave after they've gained experience. Relates to the company providing long-term incentives like income and career progression.

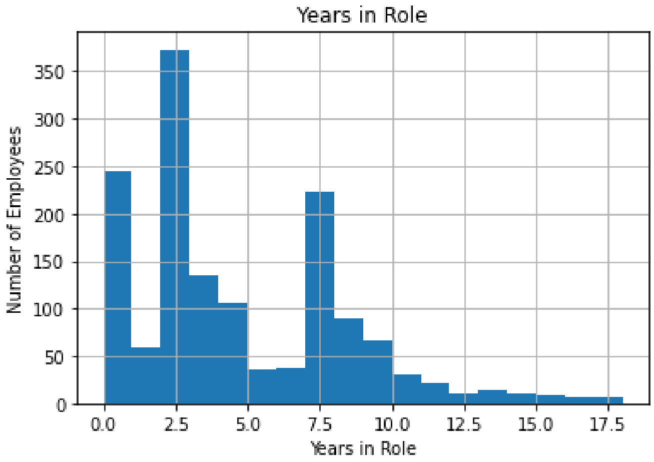
Resignation By Age: Box: depicts median and mean distribution age of current employees compared to resigned. Resignation age gives insights into the why staff are resigning.

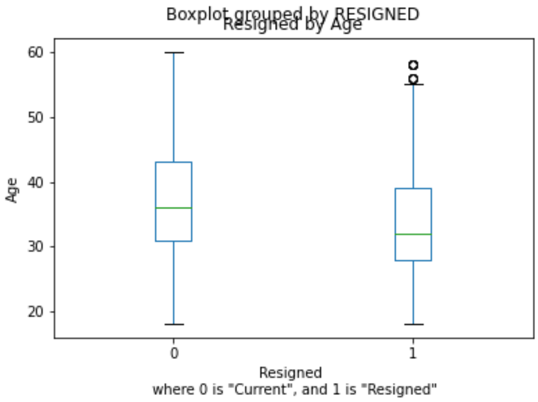
Monthly Income by Resigned Bar Graph: Created to investigate possible relationship between monthly incomes of those that resigned and current employees.

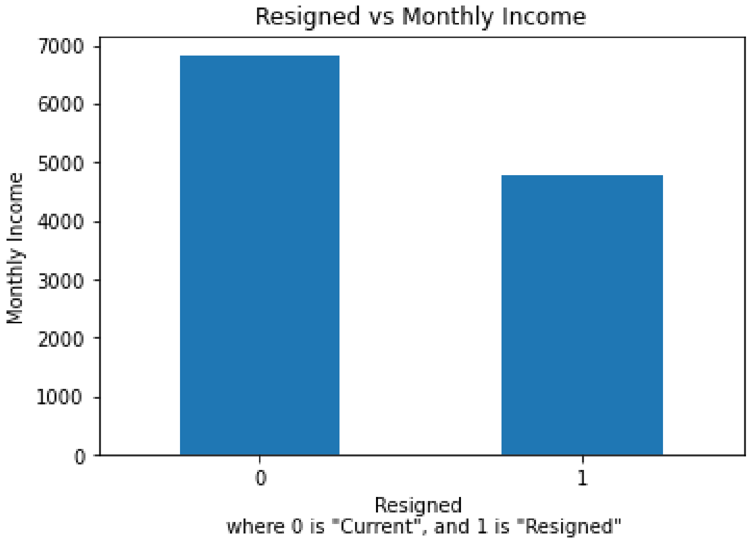
Years in Role and Job Satisfaction Bar Graph: created to investigate relationships between years spent in role and job satisfaction. It can be assumed that job satisfaction plays a role in staff retention. Are staff happy with their role? Is the company able to provide incentives, such as training and career progression to keep staff in the long-term?

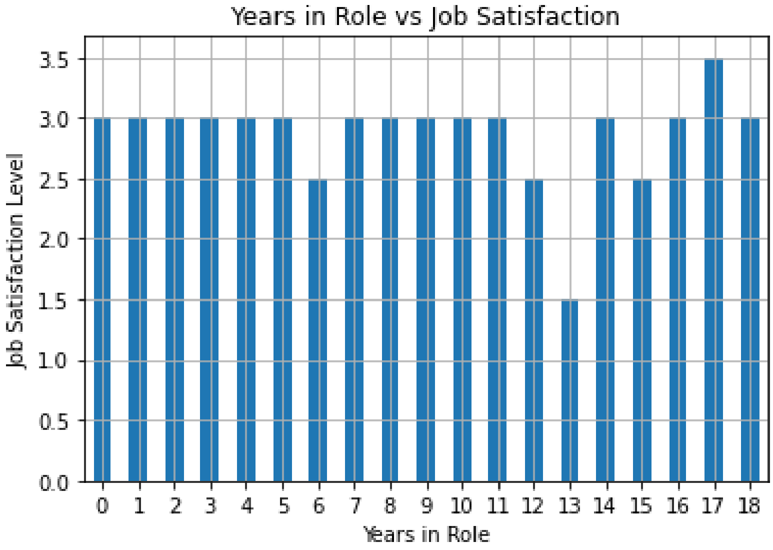












Conclusion

The data exploration that forms this report has yielded interesting relationships among the data. The aim of this report was for data preparation and initial data exploration to assist with subsequent analysis and data modelling in order to identify staff most likely to leave the company. With Issues relating to staff retention and declining quality of work, we would normally assume that these issues stem from factors including: Work life balance, Job Satisfaction, Training and income. Key factors discovered in our initial data exploration is that the distribution of years employees have spent in their roles is largely short, with majority of employees having spent four years in the company. Average age of employees resigning is much lower than those currently employed. This also applies to monthly incomes of those resigning being much lower on average compared to current employees. Another interesting observation is that there is no relationship between job satisfaction when compared to years employees have spent in their role. Early observations in this report support early hypothesis that Revolution Consulting is losing experienced and capable employees to competing firms. A by-product of which is having less experienced employees and a decline in work quality. This will assist with further analysis and data modelling in later stages.

References

Stack Overflow

Invernizzi, L 2015, How to drop a list of rows from Pandas dataframe?, Stack Overflow, viewed 21 Jan 2022

< https://stackoverflow.com/questions/14661701/how-to-drop-a-list-of-rows-from-pandas-dataframe/>

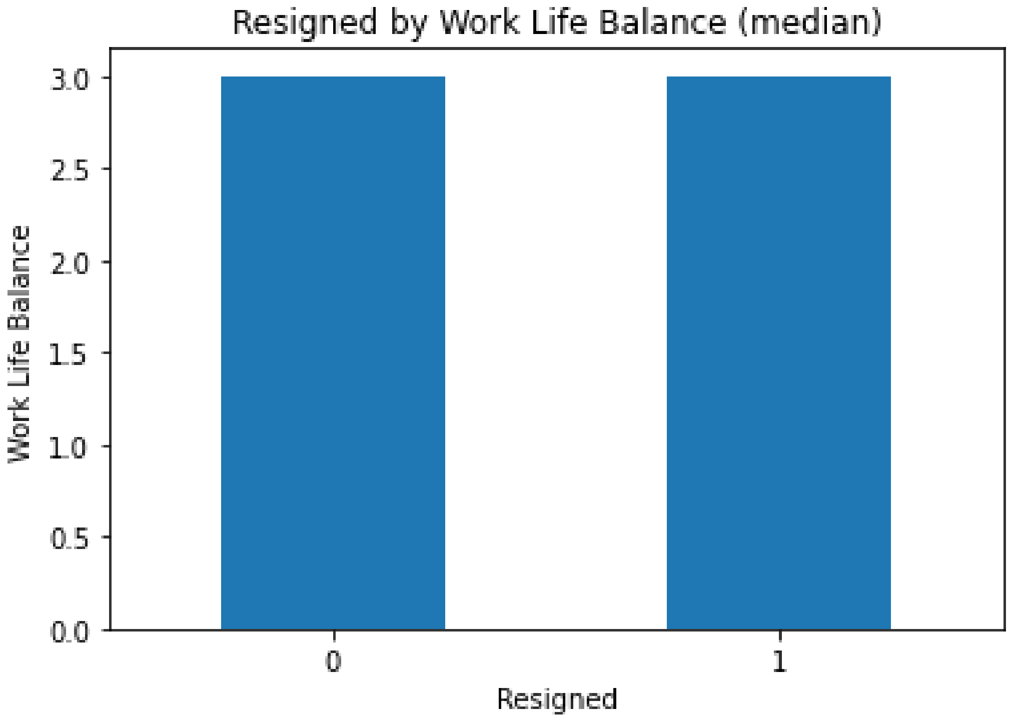
CMDlinetips.com

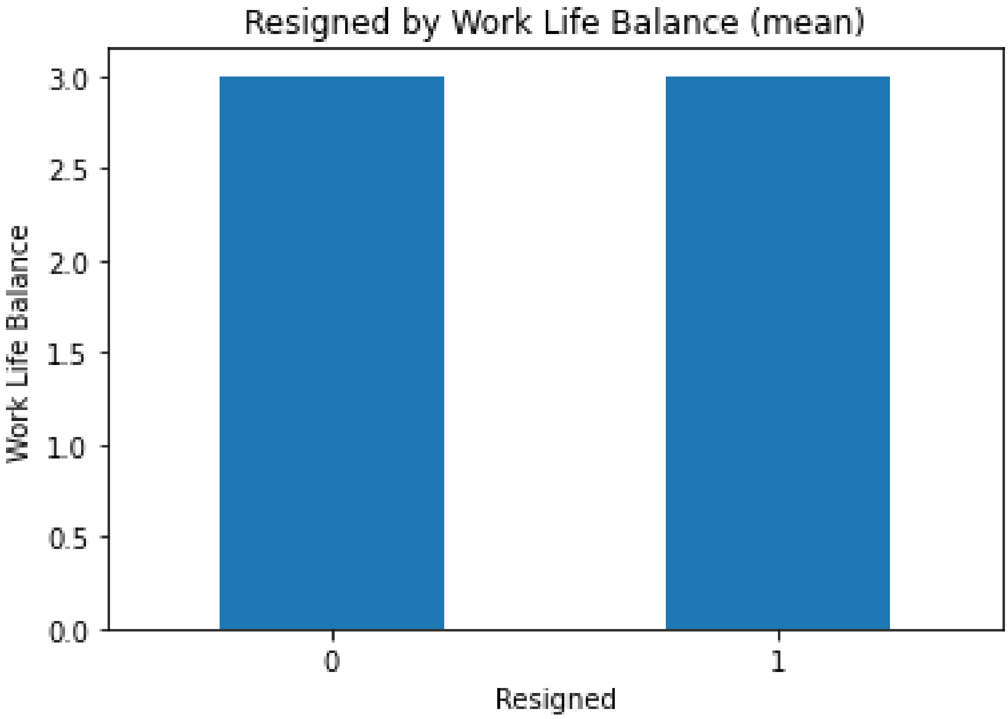
Cmdline, 2020, *How To Change Pandas Columns Names to Lower Case*, cmdlinetips.com, viewed 21 Jan 2022

<https://cmdlinetips.com/2020/07/cleaning_up_pandas-column-names/>

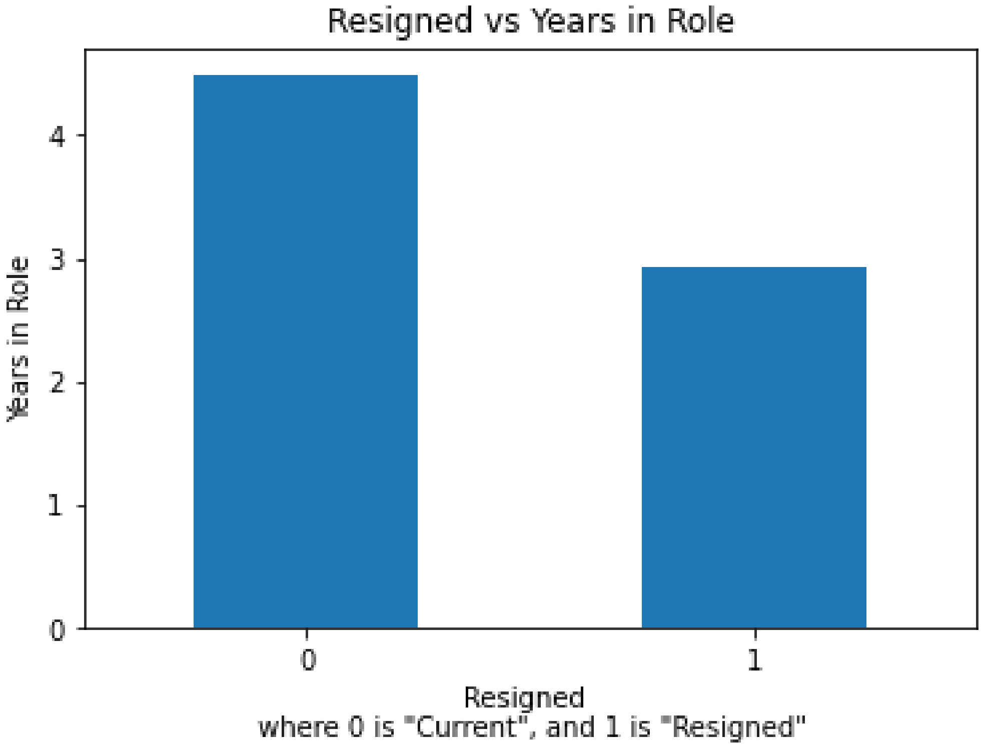
Appendix

Appendix 1: *Work Life Balance by Resigned (mean and median). The graph depicts no correlation between current employees and resigned employees when compared to Work Life Balance.*





Appendix 2: *Resigned by Years in Role shows mean for employees that have resigned is 2.93 years, compared with 4.48 years of current employees.*


Resigned vs Years in Role
Years in Role
Resigned
where 0 is "Current", and 1 is "Resigned"