

Assessment 1: Dataset preparation report

Aliia Gismatullina s4051304

Introduction

This report addresses data quality issues at Revolution Consulting, a fictional IT consulting company. Its primary goal is to prepare the data for subsequent analysis, specifically focusing on identifying employees at risk of leaving the organization. This report is the first step in a data science project aimed at mitigating employee attrition (RMIT University, School of Science, 2024).

In the data science pipeline, this report falls under the data preparation and data exploration phase, ensuring clean, reliable data for analysis. I will explore the case study, highlighting Revolution Consulting's challenges – declining work quality and increased employee turnover. Data science plays a pivotal role in tackling these issues.

Throughout this report, I will identify and address data quality problems, including missing fields, incorrect data types, spelling errors, and outliers. These issues must be resolved to ensure accurate analyses.

Data Preparation

Overview

The data preparation process was carried out to ensure that the dataset is in a suitable format for further analysis and modeling. The steps taken aimed to standardize data types, handle missing values and convert categorical variables into appropriate formats for analysis. These steps contribute to a cleaner and more structured dataset for data science tasks.

Process

1.1. Loading the Data

- The dataset has been loaded from the provided CSV file.
- Specified the 0th column as an index column for better data handling.
- The first five rows of the DataFrame were displayed to ensure the data was loaded correctly.

1.2 Data Curation

In this section, I performed various data curation steps to prepare the dataset for analysis.

- 1.2.1 **EmployeeID** has been specified as an index column while loading the data. To improve the index column, which does not serve as a feature during machine learning, I replaced it with new random numbers using the 'sample' function in pandas.
- 1.2.2 **Age** column has been scanned for missing values, outliers, and typing errors; converted into an integer data type.
- 1.2.3 For the **Resigned** column, I standardized the values, keeping only 'Yes' and 'No'. I also addressed missing values ('nan') by replacing them with the mode value, which is the most frequent value in the column. Finally, I converted the values to binary (1 for 'Yes' and 0 for 'No').
- 1.2.4 I reviewed the **BusinessTravel** column and standardized the values by converting them to uppercase. To simplify the categories, I removed the word "travel" from the values and converted them to categorical data types: 'RARELY', 'FREQUENTLY', and 'NON_TRAVEL'.
- 1.2.5 For the **BusinessUnit** column, first, I identified a 'Female' value, which turned out to be misplaced. I found this specific row, where the 'Sales' value was under the 'Gender' column. I swapped the values, using the row index number. Then I standardized the values by converting them to uppercase. I also corrected the term 'BUSINESS OPERATIONS' to 'OPERATIONS' for consistency and converted the column to a categorical data type with three categories: 'CONSULTANTS', 'OPERATIONS', and 'SALES'.
- 1.2.6 I addressed the **EducationLevel** column by identifying and handling missing values ('nan') by replacing them with the mode value. Then, I converted the float values to ordinal categories with the order [1, 2, 3, 4, 5].
- 1.2.7 For the **Gender** column, I standardized the values by removing white spaces, correcting spelling mistakes ('MMale' to 'Male'), converting all values to uppercase, replacing 'MALE' with 'M' and 'FEMALE' with 'F', and finally converting the column to a categorical data type with categories 'M' and 'F'.
- 1.2.8 I reviewed the **JobSatisfaction** column, identified missing values ('nan'), and replaced them with the mode value. Then, I converted the float values to ordinal categories with the order [1, 2, 3, 4, 5].

- 1.2.9 For the **MaritalStatus** column, I standardized the values by removing white spaces, replacing full words with their first letters ('Married' to 'M', 'Divorced' to 'D', 'Single' to 'S'), and converting the column to a categorical data type with categories 'D', 'M', and 'S'.
- 1.2.10 I addressed the **MonthlyIncome** column by identifying and handling missing values using the median value for imputation. I chose median value over mode and mean, as the data distribution on the histogram appeared to be skewed. Then, I converted the column to an integer data type for convenience. No significant outliers were detected.
- 1.2.11 The **NumCompaniesWorked** column did not have missing values and contained integer values from 0 to 9. No significant outliers were detected.
- 1.2.12 **OverTime** column had 3 missing values; upon visualization, I decided to replace them with the mode value. Similar to Resigned, I converted the values to binary, where 'Yes' is 1 and 'No' is 0.
- 1.2.13 In the **PercentSalaryHike** column I noticed a rare occurrence of zero in the data. To gain a better understanding of its distribution, I visualized it using a histogram (see Appendix 1.2). During this exploration, I identified a single instance with a zero value. With 5 years at the company and 4 years in the current role, this employee has never received a promotion, which is evident from the 'years since last promotion' being 0. Additionally, their performance rating is the lowest (2). Considering these factors, it's reasonable to conclude that this employee has never received a salary increase, which proves that this zero value is not an outlier.
- 1.2.14 To assess the **PerformanceRating** column, I first checked its unique values and then proceeded to visualize its distribution through a histogram. This examination revealed that there were no missing values or outliers, indicating that the data in this column was complete and free from anomalies.
- 1.2.15 I investigated the **AverageWeeklyHoursWorked** and identified an outlier with a value of 400 hours. To handle this outlier, I replaced it with the median value within a reasonable range of 40 to 71 hours. This adjustment aimed to ensure that the dataset remained consistent and free from extreme values. Subsequently, I converted the float to an integer data type for easier calculation. I created a histogram to visualize the distribution of this column after the outlier handling (see Appendix 1.3).
- 1.2.16 I examined **TrainingTimesLastYear**, **TotalWorkingYears**, **YearsAtCompany**, **YearsInRole**, **YearsSinceLastPromotion**, and **YearsWithCurrManager** columns by reviewing their unique values and visualizing their distribution through a histogram. This exploration revealed that there were no missing values or outliers in the dataset. The data in these columns appeared to be reliable and consistent for analysis.
- 1.2.17 For the **WorkLifeBalance** column, I began by checking its unique values and identified one missing value. To address this, I replaced the missing value with the mode (the most frequent value) to maintain data integrity. Additionally, I converted the float values to ordered categories, ranging from 1 (lowest) to 5 (highest), to make the data more interpretable.

Issues discovered

Before commencing the data preparation process, I developed several functions to automate tasks and avoid repetitive coding.

#	Issue name	Location	Code to identify	Rationale and solution
1	'36a' typing error	Age column	<code>'get_sorted_unique(df, 'Age')</code> function	It appears to be a typographical error, so I corrected it by replacing '36a' with '36'.
2	'Object' data type	Age column	<code>'get_sorted_unique(df, 'Age')</code> function	The typing error likely led to this variable being read as an object. Converted to an integer.
3	Inconsistency of values	Resigned column	<code>df['Resigned'].unique()</code>	1. Replaced with consistent 'Yes' and 'No'; 2. Mapped to binary (0 or 1) values for uniformity.
3	Missing values	Resigned column	<code>'count_missing_values'(df, 'Resigned')</code> function	Replaced with the mode 'No' (the most frequent value) to maintain data integrity.
4	Inconsistency of values	BusinessTravel column	<code>df['BusinessTravel'].unique()</code>	1. Converted to uppercase; 2. Replaced the values, removing 'travel'; 3. Converted to a category.
5	Inconsistency of values	BusinessUnit column	<code>df['BusinessUnit'].unique()</code>	1. Handled an inappropriate value 'Female' – swapped with the Gender column; 2. Uppercased the values; 3. Formatted the writing - 'Business Operations' to 'Operations'; 4. Converted to a category.

6	Missing values	EducationLevel column; JobSatisfaction column; OverTime column WorkLifeBalance column	<code>count_missing_values(df, column_name)</code> function	A single missing value was identified and replaced with the mode (the most frequent) value.
7	Incorrect Data type	EducationLevel column; JobSatisfaction column	<code>df['EducationLevel'].dtype</code> <code>df['JobSatisfaction'].dtype</code>	Converted floats to ordinal data types, where 1 is the lowest and 5 is the highest level.
8	Inconsistency of values	Gender column	<code>df['Gender'].unique()</code>	1. Removed the white spaces; 2. Corrected the spelling mistakes - 'MMale'; 3. Bring all values to uppercase; 4. Replace the 'MALE' with 'M' and 'FEMALE' with 'F'; 5. Convert an object to a category data type.
9	Inconsistency of values	MaritalStatus column	<code>df['MaritalStatus'].unique()</code>	1. Removed whitespaces; 2. Replaced the full words with their first letters. 3. Converted to a category.
10	Missing values	MonthlyIncome column	<code>count_missing_values(df, 'MonthlyIncome')</code> function	1. Calculated mean, median, mode. 2. When all 3 values appeared to differ, created a histogram to visually see the data distribution. 3. It was right-skewed, so I chose the median to replace the 3 missing values. (see Appendix 1.1)
11	Outlier	AverageWeeklyHoursWorked column	<code>'get_sorted_unique(df, column_name)'</code> function	1. Defined a reasonable range for weekly hours (40 to 71). 2. Calculated the replacement value (median within the range). 3. Replaced the outlier with the replacement value.

Data Exploration

Overview

In this exploratory data analysis, I focused on selecting and analyzing three features from the dataset. These columns were chosen to represent different data types: nominal, ordinal, and interval/ratio. The selected columns are as follows:

- Nominal: **BusinessTravel**;
- Ordinal: **EducationLevel**, **JobSatisfaction**, **PerformanceRating**, **WorkLifeBalance**
- Interval/Ratio: **Age**, **MonthlyIncome**, **YearsAtCompany**.

I aimed to understand the distribution of each selected column and explore potential relationships between columns to gain insights into the dataset.

Process

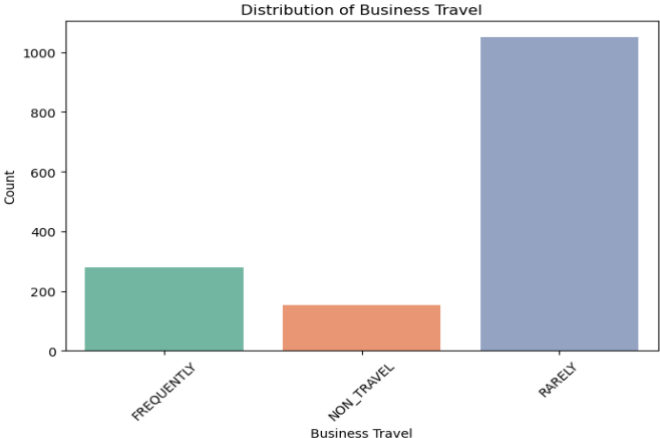
1. **BusinessTravel** variable (Bar Chart)
 - I chose a bar chart to visualize the distribution of different Business Travel frequencies among employees.
 - Justification: A bar chart is suitable for showing the frequency distribution of a nominal variable.
 - Observation: Most employees rarely travel for business, followed by those who travel frequently, while a smaller portion does not travel for business at all.
2. **JobSatisfaction** variable (Box Plot)
 - I used a box plot to examine the distribution of Job Satisfaction levels among employees.
 - Justification: A box plot is useful for visualizing the distribution of ordinal data and identifying potential outliers.
 - Observation: Job Satisfaction levels are fairly evenly distributed across the range of Monthly Income, and there don't appear to be significant outliers.
3. **Age** variable (Histogram)
 - I created a histogram to visualize the distribution of employee ages in the dataset.
 - Justification: A histogram is suitable for showing the distribution of interval/ratio data.
 - Observation: Employee ages are relatively normally distributed, with a peak in the mid-30s to mid-40s age range.
4. Exploring relationships – **EducationLevel** and **MonthlyIncome** (Box Plot)
 - I examined the relationship between an employee's Education Level and their Monthly Income using a box plot.

- Justification: A box plot helps compare the distribution of one variable across different categories of another variable.
 - Observation: There is a trend of higher Monthly Income for employees with higher Education Levels, suggesting a positive relationship.
5. Exploring relationships – **JobSatisfaction** and **WorkLifeBalance** (Bar Chart)
- I investigated the relationship between Job Satisfaction and Work-Life Balance using a bar chart.
 - Justification: A bar chart is suitable for exploring relationships between two ordered categorical variables.
 - Observation: The bar chart clearly illustrates that as Work-Life balance increases, Job Satisfaction tends to improve. This observation highlights a positive correlation between an employee's satisfaction with their job and their perception of work-life balance.
6. Exploring relationships – **Age** and **YearsAtCompany** (Scatter Plot)
- I explored the relationship between an employee's Age and the number of years they have spent at the company using a scatter plot.
 - Justification: A scatter plot is appropriate for visualizing the relationship between two continuous variables.
 - Observation: There is a somewhat positive relationship between Age and Years at Company, indicating that older employees tend to have more years of experience at the company.

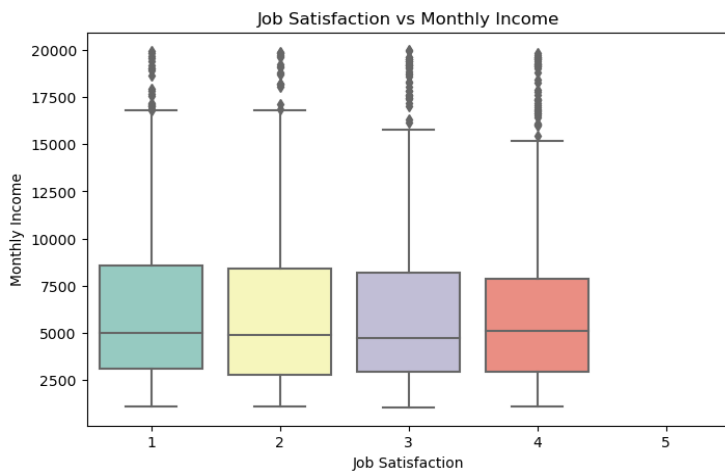
Observations

#	Observations	Significance
1	BusinessTravel distribution: ○ Most employees rarely travel for business.	Understanding the distribution of business travel frequencies can help in workforce planning and travel expense management.
2	JobSatisfaction levels: ○ Fairly evenly distributed across MonthlyIncome.	This observation suggests that Monthly Income does not strongly correlate with Job Satisfaction, which could inform HR policies on compensation and employee satisfaction.
3	Age Distribution: ○ Employee ages are relatively normally distributed.	Knowing the age distribution can be crucial for workforce planning and tailoring HR practices to different age groups.
4	EducationLevel and WorkLifeBalance : ○ Higher education levels tend to be associated with higher Monthly Income.	Recognizing this relationship can assist in compensation and career development strategies.
5	JobSatisfaction and WorkLifeBalance : ○ A positive correlation is evident. Higher Job Satisfaction aligns with better Work-Life Balance ratings, while lower Job Satisfaction corresponds to lower Work-Life Balance scores.	This correlation highlights the importance of simultaneously improving Job Satisfaction and Work-Life Balance to enhance the overall work experience. Happy employees tend to perceive better Work-Life Balance, promoting productivity and retention.
6	Age and YearsAtCompany : ○ Older employees tend to have more years of experience at the company.	This observation highlights the experience-age relationship, which can influence employee retention and career progression strategies.

Plots

Plots	Observations
1. BusinessTravel (Bar Chart)	
 <p>The bar chart displays the distribution of business travel frequencies among employees. The 'RARELY' category has the highest count, exceeding 1000. The 'FREQUENTLY' category has a count of approximately 280, and the 'NON_TRAVEL' category has the lowest count at approximately 150.</p>	<p>Question: How is BusinessTravel distributed among employees?</p> <p>Observation: Most employees rarely travel for business, followed by those who travel frequently. Only a small number of employees don't travel for business at all. This provides an overview of the Business Travel patterns among employees.</p>

2. JobSatisfaction (Box Plot)



Question: Does **MonthlyIncome** exhibit a significant influence on **JobSatisfaction** levels?

Observation: **JobSatisfaction** levels appear to be fairly evenly distributed across different **Monthly Income** ranges. This observation suggests that **Monthly Income** does not strongly correlate with **Job Satisfaction**, indicating that other factors may play a more substantial role in determining employees' satisfaction with their jobs.

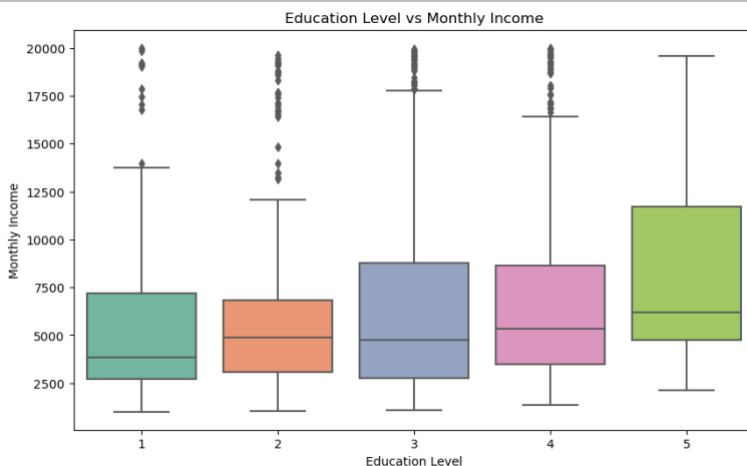
3. Age (Histogram)



Question: What is the distribution of employee **Ages**?

Observation: The histogram indicates that the majority of employees are in the range of approximately 30 to 40 years old, with a relatively even distribution within that range. There is a decrease in the number of employees in the older age groups, which is expected.

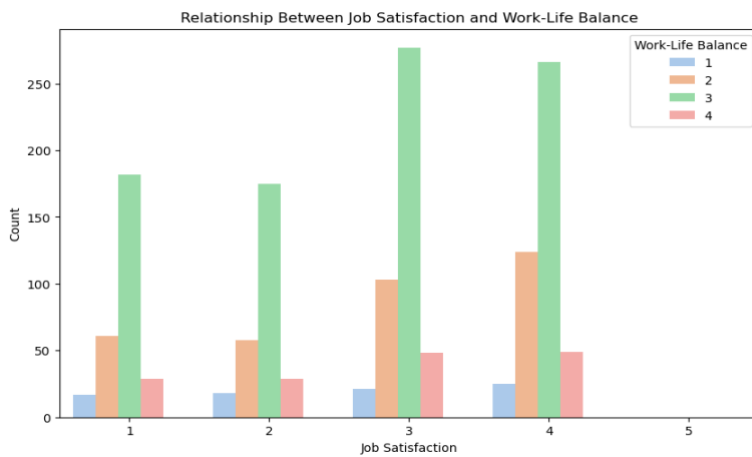
4. EducationLevel and MonthlyIncome (Box Plot)



Question: Is there a correlation between an employee's **EducationLevel** and their **MonthlyIncome**?

Observation: It appears that higher education levels are associated with higher **Monthly Incomes**. This observation suggests that employees with higher educational qualifications tend to earn more. Recognizing this relationship can be valuable for designing compensation and career development strategies that take into account an employee's educational background when determining income levels.

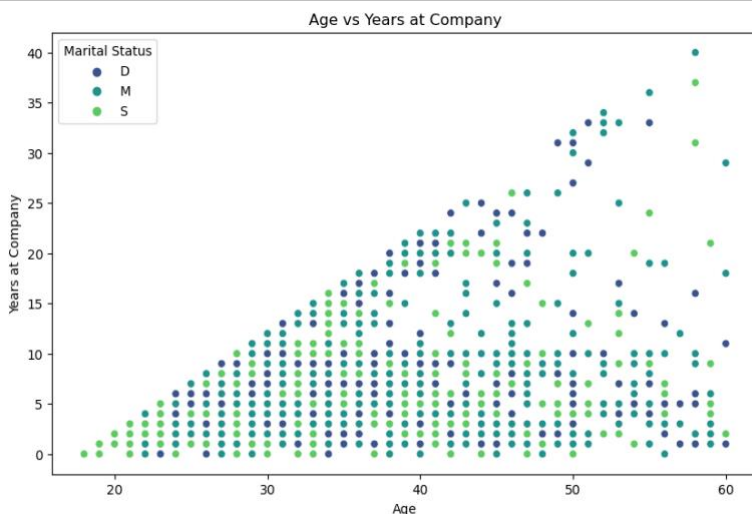
5. JobSatisfaction and WorkLifeBalance (Bar Chart)



Question: Is there a relationship between Job Satisfaction and Work-Life Balance?

Observation: The bar chart clearly shows that as **WorkLifeBalance** increases, **JobSatisfaction** tends to improve. This suggests a positive correlation between an employee's satisfaction with their job and their perception of work-life balance.

6. Age and YearsAtCompany (Scatter Plot)



Question: Is there a correlation between an employee's **Age** and the number of years they have worked at the company?

Observation: It's evident that older employees tend to have more years of experience at the company. This observation highlights the positive correlation between age and the duration of employment at the company. Understanding this relationship is crucial for HR strategies related to employee retention and career progression, as it underscores the significance of experience in an employee's tenure at the company.

Conclusion

In conclusion, this report has successfully addressed data quality issues at Revolution Consulting, paving the way for further analysis to identify employees at risk of leaving the organization. The data preparation phase involved standardizing data types, handling missing values, outliers, typing errors, inconsistent variables, and converting variables into appropriate formats for analysis. These steps ensured a clean, reliable dataset, setting the stage for effective data science tasks.

Key observations:

- **EducationLevel and MonthlyIncome:** higher education levels tend to be associated with higher monthly income. Recognizing this relationship can assist in compensation and career development strategies.
- **JobSatisfaction and WorkLifeBalance:** improved work-life balance correlates with higher Job Satisfaction, emphasizing the importance of balancing employee satisfaction and work-life.
- **Age and YearsAtCompany:** older employees tend to have more experience at the company, offering insights into employee retention and career progression.

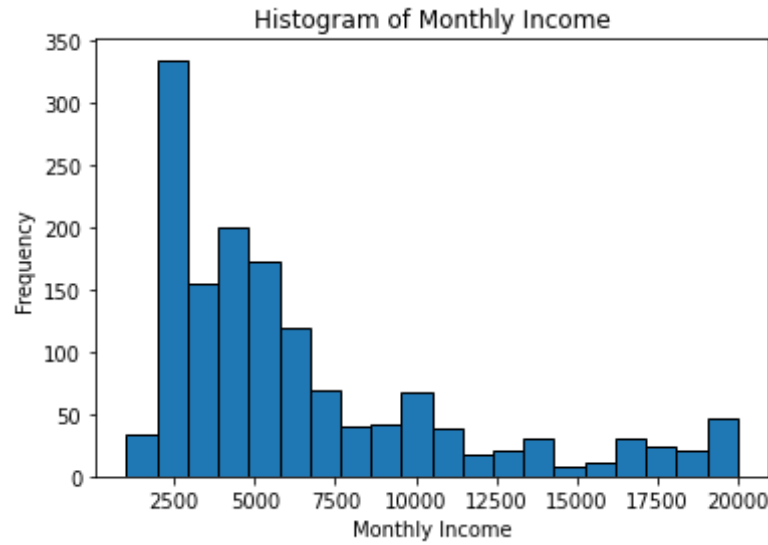
These observations will contribute to addressing attrition issues and aligning with the company's objectives.

References

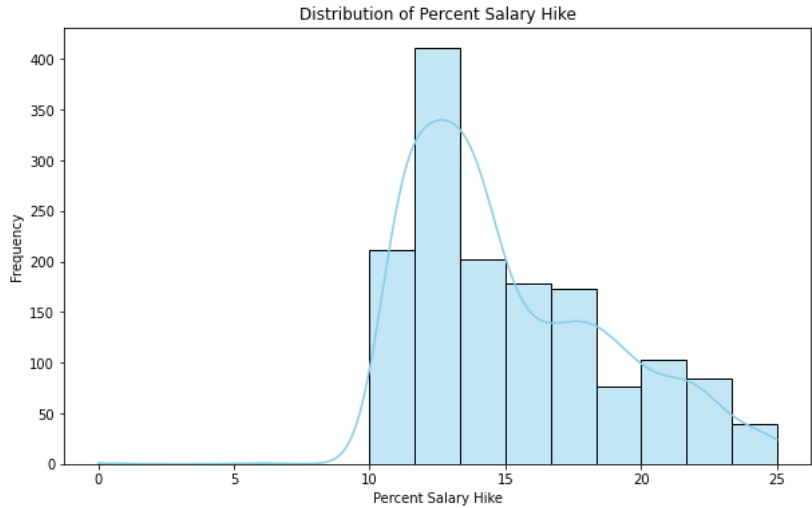
RMIT University, School of Science. (2024). Assessment 1 Case Study.
<https://rmit.instructure.com/courses/128572/files/35200366?wrap=1>.

Appendix 1: Data Preparation Visualizations

Appendix 1.1: Histogram of Monthly Income



Appendix 1.2: Distribution of Percent Salary Hike



Appendix 1.3: Distribution of Average Weekly Hours Worked

