Assessment 1: Dataset preparation report

1. Introduction

This report is to conduct cleaning and exploration of employee data provided by the HR department of Revolution Consulting. It is a critical step of the data science pipeline, ensuring the data is tidy, accurate and free of errors. Properly cleaned data is the foundation of a meaningful analysis. The goal is to prepare the data for deeper analysis that will help identify factors to employee turnover.

2. Data preparation

2.1 Overview

The dataset helps identifying patterns why Revolution Consulting is experiencing high employee turnover. It lists information about employee demographics, job satisfaction, work-life balance, and other factors that may influence consultant's decision to stay or leave the company. **Unchanged** Features are:

Object Type Variables	**Unique values**	**Object Type Variables**	**Unique values**
Age	44	Resigned	6
BusinessTravel	6	BusinessUnit	4
Gender	8	MaritalStatus	6
Ovetime	2		
Float64 Type Variables	**Unique values**	**Float64 Type Variables**	**Unique values**
EducationLevel	5	JobSatisfaction	4
MonthlyIncome	1354	AverageWeeklyHoursWorked	24
WorkLifeBalance	4		
Int64 Type Variables	**Unique values**	**Int64 Type Variables**	**Unique values**
EmployeeID	1482	NumCompaniesWorked	10
PercentSalaryHike	18	PerformanceRating	3
TotalWorkingYears	40	TrainingTimesLastYear	7
YearsAtCompany	37	YearsInRole	19
YearsSinceLastPromotion	16	YearsWithCurrManager	18

2.2 Process

1. We import all necessary libraries for this Dataset into Jupyter Notebook, such as Pandas, Numpy, Matplotlib and Seaborn
2. We load the csv - file and display the rows and columns of our dataframe. We perform some descriptive statistics on our data. The pandas function pd.set_option() makes sure all columns are being printed in the following .head() function *(Pandas.pydata.org, 2024)*.
3. Next, we check the data type of all variables using the method .dtypes but we don't change the types yet, as the data is not tidy.
4. We check our data for Typos and verify the results using .value_counts() on 'Age', 'Resigned', 'BusinessTravel', 'Gender', and 'MaritalStatus'. Finally, we fix the typos using .replace().
5. The output of our for loop also revealed some whitespace in every column. We apply a for loop that applies a lambda function, which checks if the value is a string. If it is the function .strip() will be applied which removes any extra spaces at the beginning or end of each string *(GeeksforGeeks, 2017)*.
6. Next, we change all string values to upper-case letter.

7. We split the Sanity Check in two parts. The following first one will check if all qualitative values are realistic. We see that the columns 'BusinessUnit' includes a wrong entry 'Female' and 'Gender' the value 'Sales'. We apply np.nan to change the wrong entries to missing values *(Numpy.org, 2024)*.
8. To display all missing values, we call the function .isna().sum() on our dataframe. Next, we Subset all character values in a list and iterate over it using a for loop, which replaces missing values with the column specific mode char_mode. We follow the same procedure to replace all missing values for our numeric columns, but instead of using the mode we use the mean num_mean in our .fillna() function *(Pandas.pydata.org, 2024)*.
9. Next, we display all columns and their data types using .info(). Using the .astype() function, we ensure that every column has the correct data type. 'EmplyoeeID' should be an object to avoid data leakage, Age needs to be numerical and all ordinal data needs gets transformed to a categorical data type. We also change Resigned and Overtime to a Boolean as they display two values, YES and NO.
10. Since all columns are tidy and in their correct data type, we perform a sanity check on our numerical columns using .describe(). The column 'AverageWeeklyHoursWorked' has an unrealistic entry '400' hours in a week, we display the variable in a boxplot to visualize its distribution. We replace the columns max value with its median – median to disregard the outlier's effect on the columns mean.

2.3 Issues discovered

The table below to consists of all issues and their fixes.

#	Issue name	Location	Code to identify	Rationale and solution
1	Fixing typos	Typos found in column 'Age', 'Resigned', 'BusinessTravel', Gender' and 'MaritalStatus'	Performed df.value_counts() on the columns to display all unique values, discovered typo such as '36a'; 'Y','N'; 'Travels_Rarely','rarely_travel'; 'MMale','M'; 'D'.	Used the function df.replace(old_value, new_value) to fix the typo. df['Gender'].replace(['MMale', 'M'], 'Male', inplace = True)
2	Fixing whitespace	Whitespace found in the string values	Using a for loop to print all columns and their unique values .unique. '\n' helps to display the variable illegible. <Code next cell below>	Performing x.strip() on string values to delete whitespace of all quantitative variables. The lambda function skips numerical values as they can't contain whitespace.
3	Converting to Upper-case	Multiple entries of the same value but using lower and upper cases.	Using the previous for loop to display all unique value. column_names = df.columns for column in column_names: print(f"Unique values in {column}:") print(f"{df[column].unique()}\n")	Subsetting all string variables to apply the string function .str.upper(). df[categorical_data.columns] = categorical_data.apply(lambda x:x.str.upper())
4	Converting nonsense qualitative values to NaN	Wrong Data found in the columns 'BusinessUnits' and 'Gender'.	Displaying all column specific values using the function .unique on every column using a for loop. Gender has a wrong entry 'SALES' and BusinessUnits 'FEMALE'. df.value.counts() to verify wrong entries.	Using the numpy function np.nan to convert the wrong entries to missing values, so we can replace them with their mode value later. df.loc[df['Gender'] == 'SALES', 'Gender'] = np.nan

5	Fixing missing values in qualitative variables	Missing qualitative values in the dataframe	Using .isna().sum() on our dataframe df to display all columns with their total missing data. df.isna().any(axis=1)	Subsetting all quantitative columns as a list to apply a for loop, which iterates over the list and replaces the missing values NaN of each column with their mode using .fillna().
6	Fixing missing values in ratio variable	Missing value in ratio column 'MonthlyIncome'	Using df.isna().sum() on our dataframe df to display all columns with their total missing data.	Using .fillna() to replace NaN in 'MonthlyIncome' with the columns mean value.
7	Correcting qualitative variable types	Changing the numeric column type EmployeeID to an object	Using df.info() on our dataframe df to display each column with their corresponding data type.	Using .astype('object') and a dataframe mask to change the data type to an object to reduce data leakage. df['EmployeeID'] = df['EmployeeID'].astype('object')
8	Correcting quantitative variable types	Changing the object column type Age and **all ordinal columns** to an integer	Using df.info() to display each column with their corresponding data type.	Using .astype('int64') and a dataframe mask to change the ordinal data to an integer, e.g: df['Age'] = df['Age'].astype('int64')
9	Fixing unrealistic values for numerical data	Outlier detected in 'AverageWeeklyHoursWorked'	Using df.describe() to detect any unrealistic numerical values in our data. sns.boxplot(data = df, y ='AverageWeeklyHoursWorked')	Using .replace() to replace the maximum of 'AverageWeeklyHoursWorked' with its median value. **Median**, to ignore the outlier.

3. Data exploration

3.1 Overview.

The exploration focuses on key features within the dataset to analyze correlations and patterns contributing to employee turnover. The visualizations cover demographics, job satisfaction, salary distributions, to provide insights how Revolution Consulting can improve their work environment. The below delves deeper into areas of inequality or motivation that impact consultant retention.

3.2 Process

1. First, we subset the columns 'BusinessUnit', 'EducationLevel' and 'MonthlyIncome'
2. We display the total count of each Business Unit of our company using a countplot, as are visualizing nominal data.
3. A boxplot displays simple statistics such as the mean, min, max, 25% range, 75% range and IQR. It is being used to display nominal data *(Seaborn.pydata.org, 2024)*.
4. To visualize the distribution of ratio data we use a Distribution plot for 'MonthlyIncome'
5. Next, we analyze relationships between variables, we start with a barplot to emphasize the relationship between 'MonthlyIncome' and 'YearsAtCompany' colored by 'Gender'
6. We plot a scatterplot to visualize the correlation between 'Age' and 'PercentSalaryHike' colored by 'EducationLevel' *(GeeksforGeeks, 2020)*.
7. Our last visualization provides a good overview of other correlation within the dataframe by generating a heatmap *(seaborn.pydata.org, 2024)*.

Observations

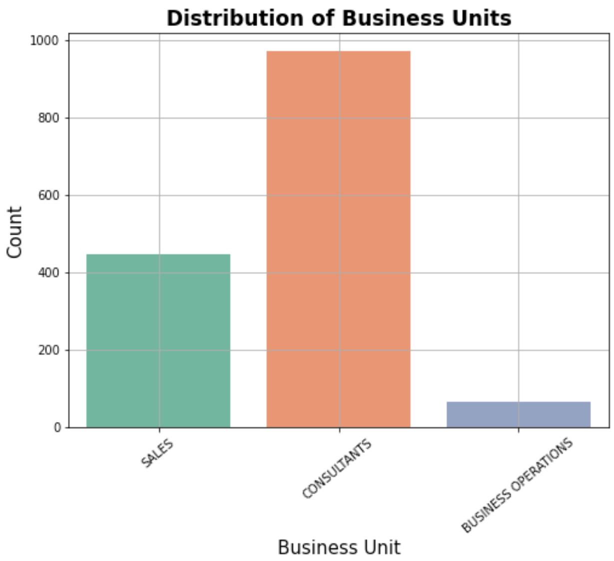
#	Observations	Significance
1	The Countplot of the nominal column BusinessUnits, shows the distribution between each of the three departments. We can see that Operations has roughly 100 members, Sales 450 and Consultants almost 1000.	The company needs enough Business Operators to ensure the quality of our consultants and to hold performance reviews to improve their work and environment. Operations might not be able to implement requests to improve the work field for the consultants due to their low number.
2	The Boxplot of Education Level by Gender shows that the mean, min, max and whiskers of the Education Level is equally distributed by Gender.	The company doesn't have any Gender dominating the Education Level, which is important to guarantee a balanced working environment.
3	The Distribution Plot is ideal to show the distribution of ratio data Monthly Income. We can see that the most employees earn approximately 3000 dollar, while the maximum salary lays around 20000 dollar.	It is important to pay the Employees equal to avoid jealousy within the company. The plot also shows potential salary opportunities for employees after a promotion to upper management. The company might need to increase their mean salary to keep their staff.
4	The Barplot 'Monthly Income by Years At Company' shows that the longer you work for our company the higher your Income category. It shows that Females earn slightly more in the first half of the Monthly Income categories and male more in the second, which major inequality within the $10k - $12k range.	The bar plot shows that our company rewards loyalty significantly and that there is a strong correlation between Monthly Income and Years At Company. This might motivate young consultants to stay loyal and grow within the business. On the other hand, the income inequality of male and female might raise concern for the consultants !
5	The Scatterplot shows the correlation between 'Age', 'SalaryPercentHike' and 'EducationLevel'. The graph emphasis that there is no correlation between the features, 'EducationLevel' and 'Age' have a slight correlation 0.2.	Our company doesn't take Age or the Education Level into consideration when discussing about a Salary Hike, the usual Hike sits between 10% and 25% depending on other factors. This result might demotivate highly qualified consultants to grow within the business!
6	The Heatplot is showing the correlation between eight different variables of all kinds to quickly identify correlations. We can see that 'Age' and 'MonthlyIncome' are moderately correlated, 'MonthlyIncome' and 'AverageWeeklyHoursWorked' are rather strong correlated.	The heatmap displays the overall culture of our company, there are some strong relations between Age, Working Years and Monthly Income. This emphasizes that our company values on the job expertise and experience more than university degrees. It might be demotivating that consultants will eventually improve their salary just due to yearly increases not performance boni.

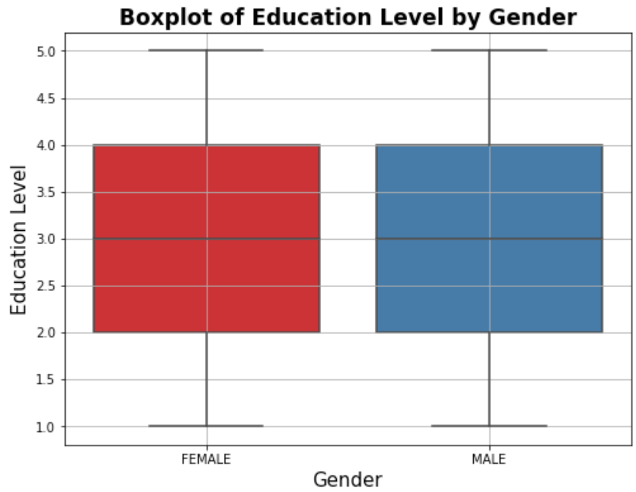
3.3 Plots

This section concentrates on the visualisation of different features and how different features correlate to each other. Every visualisation has the intention to answer a hypothesis and to provide further insights how Revolution Consulting can identify risks and drivers of employee turnover.

3.3.1 P2. Distribution of Business Units

3.3.2 P3. Boxplot of Education Level by Gender

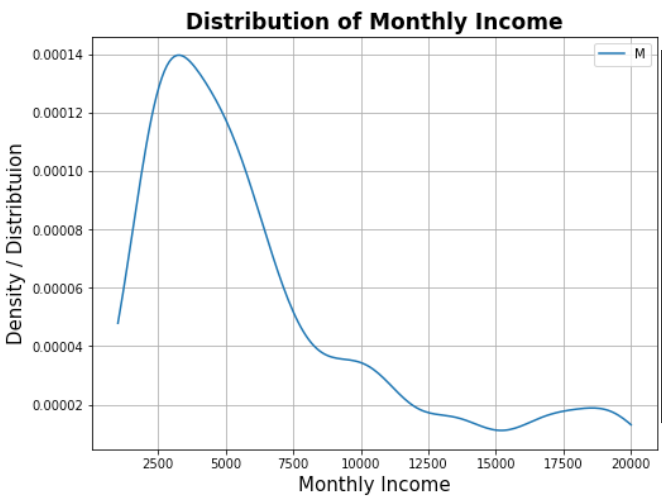




We use a Countplot to visualize the nominal column 'BusinessUnit'. This plot intends to answer the question if an imbalance of divisions could affect the company's operations? The plot proves that Revolution Consulting might need to employ more Operations to improve the environment of consultants and to assist them in their work.

The Boxplot is the perfect tool to visualize the distribution of the nominal variable EducationLevel by Gender. It answers the question if any inequality between the Education level exists and how the Education distribution of staff member look like. 75% of employee's sit between an Education Level of 2.0 and 4.0 with an average of 3.0, no inequality has been identified.

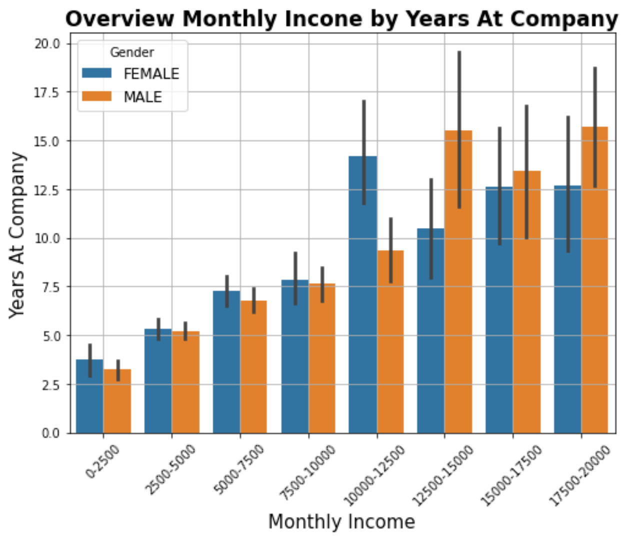
3.3.3 P4. Distribution of Monthly Income

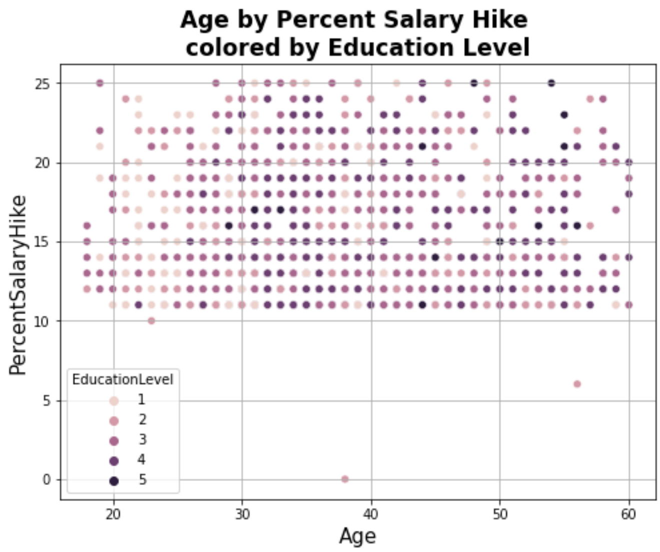


The Distribution Plot has been used to visualize the ratio data column of 'MonthlyIncome'. It answers the question if the salary of the employee's needs to be reviewed. The Distribution shows that Revolution Consulting might have a median income which lays under the market average, resulting in higher employee turnover.

3.3.4 P5. Relationship Monthly Income by Years at Company

3.3.5 P6. Age by Percent Salary Hike coloured by Education Level

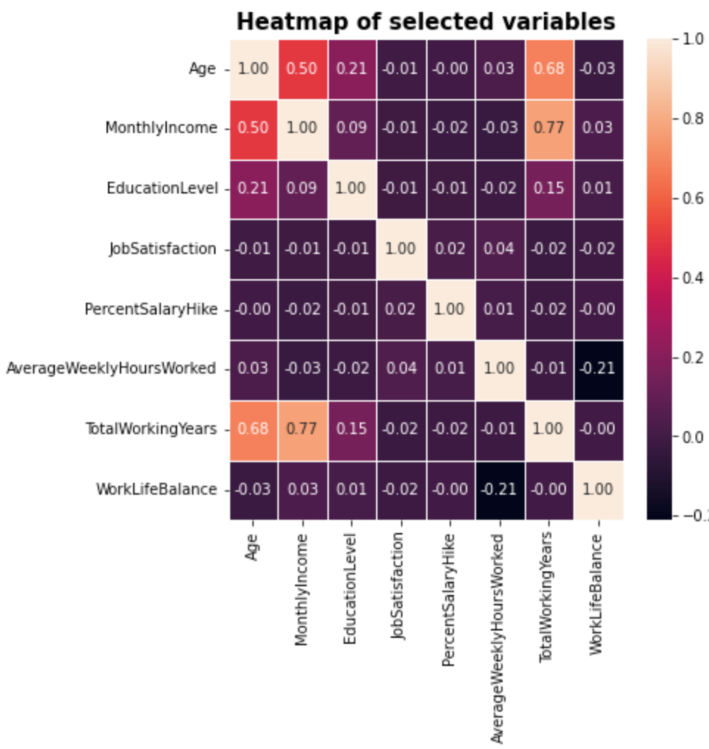




The above bar plot visualizes the relationship between 'YearsAtCompany' and 'MonthlyIncome' to answer the question if Revolution Consulting rewards loyalty towards the company financially. It also displays if a long-term carrier would pay off. The colouring by Gender emphasis if any gender-pay inequalities exist. We grouped the Monthly Income by categories for a better visualisation.

Next, the correlation between 'Age', 'PercentSalaryHike' and 'EducationLevel' shows that everyone within the company can benefit from an income increase between 10% and 25%. Unfortunately, the graph also displays the irrelevance of employee's Educations Level, higher qualifications are not being rewarded financially.

3.3.6 P7. Correlation between Eight selected Variables of All Data Kind



The last visualisation is a Heatmap as it is the best choice to generate a Correlation Overview of all ordinal variables. Its focus is to answer in what aspects Revolution Consulting needs to improve on based on high correlating features and if they influence employee turnover. We can see that the most significant correlations are feature based on how many years you have spent working and within the company. This trend suggests that the company must focus to improve the motivation of new starters who are highly educated and willing to learn. The Income should not just depend on the factor of years at the company, as this demotivated talents with new approaches the business could benefit from.

4. Conclusion

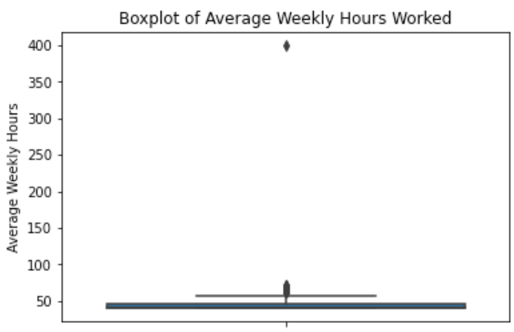
The data uncovered some key insights for Revolution Consulting, after we cleaned the inconsistencies and ensured analysis accuracy. The company needs to focus on improving the work environment for new consultants, which could be done by rewarding highly educated staff members once they solved more complicated cases. The increase of the Monthly Income is another way of keeping existing consultants. It might be necessary to even the distribution of the Business Units a little bit, so that operations run more smoothly due to more employee capacities. Revolution Consulting needs more correlating features to Monthly Income than just the Total Working Years and Age. The company also needs to further analyze the inequality of salaries between female and males, where female make more money in the first half of all income categories and males dominate the second half. The boxplot in our analysis also revealed that 75% of employees work between 40 and 47 hours per week, which is way beyond a 38 hour work week.

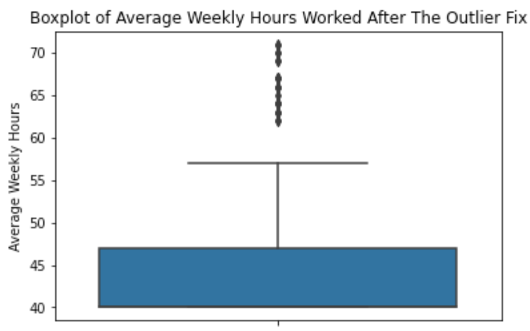
5. References

- Pandas.pydata.org. (2024). pandas.set_option — pandas 2.0.3 documentation. [online] Available at: https://pandas.pydata.org/docs/reference/api/pandas.set_option.html [Accessed 01.09.2024].
- GeeksforGeeks. (2017). Python Lambda Functions. [online] Available at: https://www.geeksforgeeks.org/python-lambda-anonymous-functions-filter-map-reduce/ [Accessed 02.09.2024].
- Numpy.org. (2024). Miscellaneous — NumPy v1.26 Manual. [online] Available at: https://numpy.org/doc/stable/user/misc.html [Accessed 05.09.2024].
- Pandas.pydata.org. (2024). Working with missing data — pandas 1.5.1 documentation. [online] Available at: https://pandas.pydata.org/docs/user_guide/missing_data.html [Accessed 06.09.2024].
- Seaborn.pydata.org. (2024). seaborn.boxplot — seaborn 0.11.1 documentation. [online] Available at: https://seaborn.pydata.org/generated/seaborn.boxplot.html [Accessed 06.09.2024].
- GeeksforGeeks. (2020). *Scatterplot using Seaborn in Python*. [online] Available at: https://www.geeksforgeeks.org/scatterplot-using-seaborn-in-python/ [Accessed 06.09.2024].
- Seaborn.pydata.org. (2024). *seaborn.heatmap — seaborn 0.10.1 documentation*. [online] Available at: https://seaborn.pydata.org/generated/seaborn.heatmap.html [Accessed 07.09.2024].

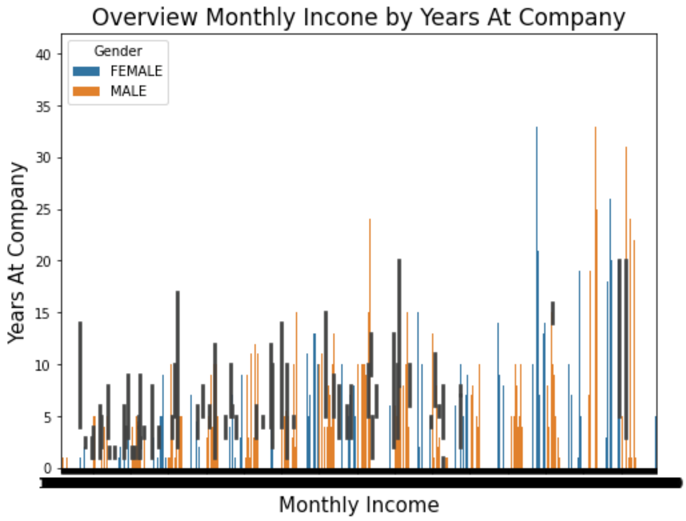
6. Appendix

6.1 Sanity Check Boxplot before and after cleaning





6.2 Overview Monthly Income by Years At Company (Uncategorized)



6.3 Cleaned dataset last five columns

	EmployeeID	Age	Resigned	BusinessTravel	BusinessUnit	EducationLevel	Gender	JobSatisfaction	MaritalStatu
1477	6680	40	True	NON-TRAVEL	CONSULTANTS	4	MALE	3	DIVORCEI
1478	3190	33	True	TRAVEL_RARELY	CONSULTANTS	4	MALE	3	SINGLI
1479	9017	38	True	TRAVEL_RARELY	CONSULTANTS	2	FEMALE	3	MARRIEI
1480	2477	32	True	TRAVEL_FREQUENTLY	SALES	4	MALE	4	SINGLI
1481	3238	36	True	NON-TRAVEL	CONSULTANTS	4	FEMALE	3	MARRIEI