Data Preprocessing

Report of the group project

2024/2025

Group **[REDACTED]**

Hubert Kołomański, **[REDACTED]**, **[REDACTED]**
Dino Meng, **[REDACTED]**, **[REDACTED]**
Guilherme Sousa, **[REDACTED]**, **[REDACTED]**

Table of Contents

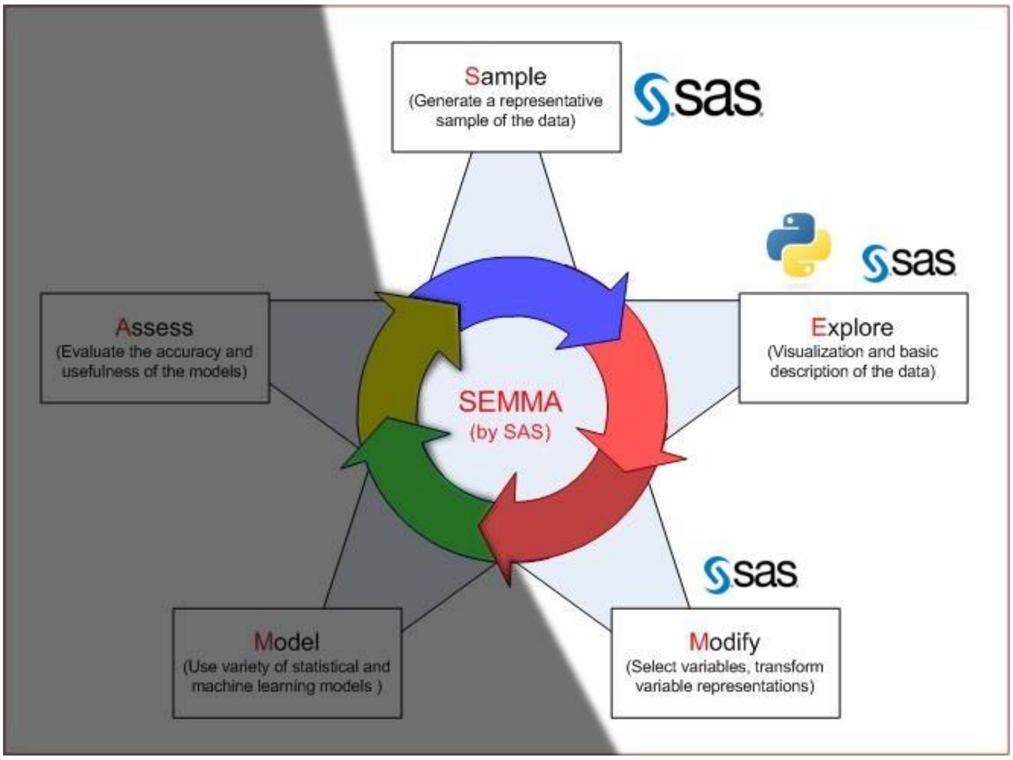
Introduction

In today's healthcare scene, patient care and satisfaction are vital. Hospitals must continuously seek ways to differentiate themselves and understand patient needs. Therefore, City Hospital, which provides services across multiple departments, aims to leverage the data collected by its information systems to enhance patient care and operational efficiency.
The data available represents patient interactions and treatments across various departments, reflecting the hospital's overall performance and patient demographics. To harness this data effectively, City Hospital's management has assembled a team of data scientists to analyze and segment patient information. Within this team, there is a dedicated subgroup focused on data preprocessing.
The data preprocessing team's role is to prepare the data for advanced analytical methods and provide initial insights into hospital operations and patient care patterns. This is crucial as the hospital currently lacks comprehensive information on its activities and patient behaviors.
City Hospital requires an exploratory analysis to address fundamental operational questions and an analytic-based table (ABT) for descriptive analysis and patient segmentation. Essentially, the DP Team aims to utilize data from the hospital's information systems to create an ABT, which will then be handed over to the next team for further analysis and implementation.

Project Methodology

As this is a *data preprocessing* project, our main pipeline concerns only the first 3 steps of the SEMMA process: Sample, Explore and Modify

- **Sample**: We will consider the transactional table a representative sample. The data will be imported with SAS Miner Enterprise.
- **Explore**: We will do exploratory data visualization on the data, to know which aspects of the dataset need particular attention. This will be mainly done with SAS Miner Enterprise, and there is minor usage of external tools such as Python.
- **Modify**: We will treat problems detected previously, mainly through two applications: SAS Miner Enterprise and SAS Guide.



(*figure 2.1.*, our partial SEMMA project pipeline)

After the data preprocessing pipeline is finished, we will build an Analytic Base Table to obtain information about the customers.

In the end, we will perform data visualizations with PowerBI to gain basic, but significant, business insights.

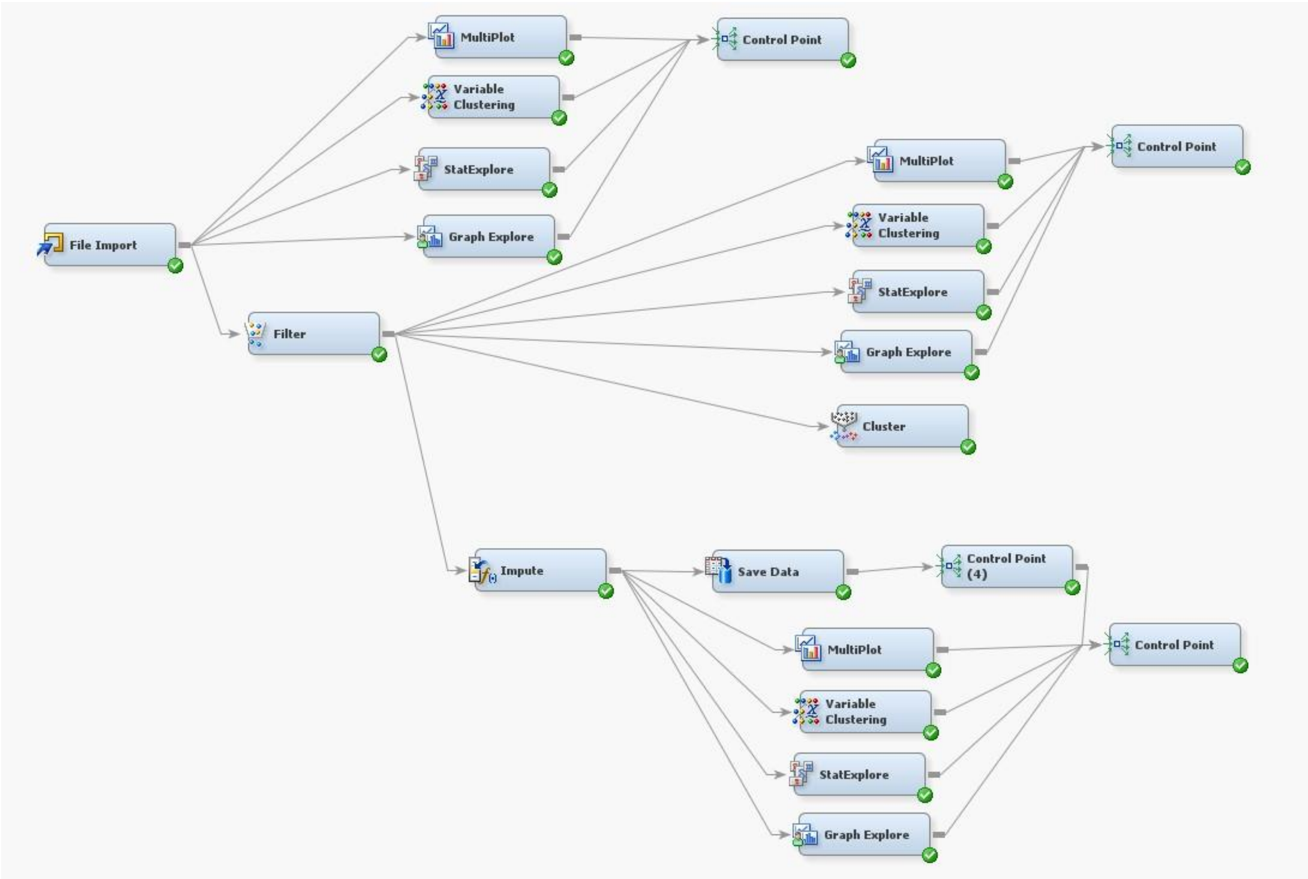
Any kind of small adjustment - such as renaming or deleting columns - will be made

on Excel. The following flow diagram (*fig. 2.2.*) represents the whole project pipeline.

(*figure 2.2.*, project pipeline)

Data Exploration and Treatment

Let us present the workflow used to explore and treat data, in SAS Miner Enterprise.



(*figure 3.0.*, SAS Enterprise Miner's diagram for the project)

Phase 0: Exploratory Data Analysis

Metadata

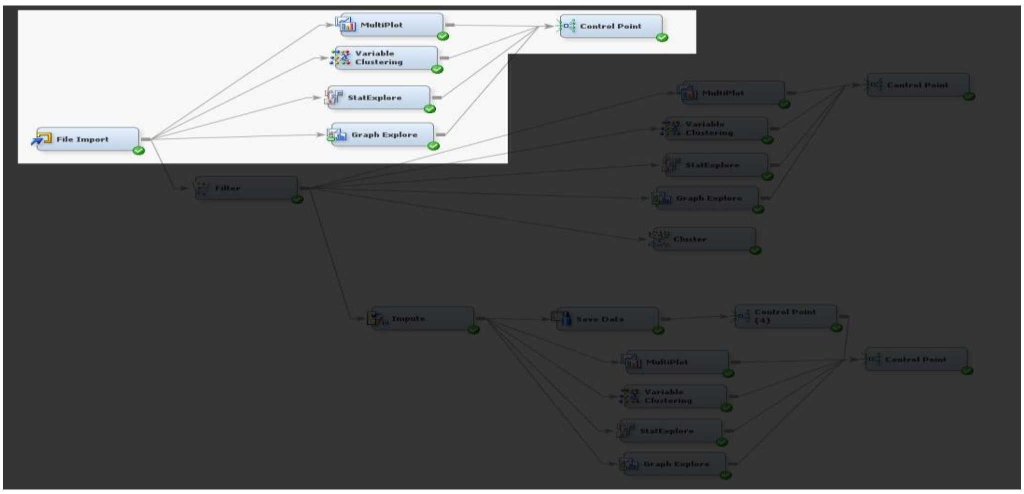
Before delving into technical details, we will explore by dataset by reading its metadata first (figure 3.1.), to gain an understanding of the business details.

Variable	Description
Patient ID	Unique identification of the patient
Age	Patient age
Gender	Patient gender (Male, Female, Other)
City of Residence	Patient city of residence
Profession	Patient profession
Insurance Provider	Patience insurance provider
Family History	Patient family history diseases
Education Level	Patient education level
Marital Status	Patient marital status
Visit Date	Date of the consultation
Department	Consultation department
Consultation Duration	Consultation duration in minutes
Satisfaction Level	Patient evaluation of the satisfaction level with the consultation (1-5)
Approximate Annual Income	Patient approximate annual income
Consultation Price	Consultation price (pounds)
Insurance Coverage	Amount of the consultation price covered by the insurance provider (pounds)

(*figure 3.1.*, metadata provided by project guidelines)

The initial dataset provided City Hospital is a transactional table containing information about each patient visit; therefore, it is crucial to ensure that each transaction has correct values, to perform clustering on the transactions and patients. The dataset contains information about 10008 transactions.

EDA With SAS Miner Enterprise



(*figure 3.2.*, EDA with SAS Miner Enterprise)

Then we performed an initial inspection of the dataset through SAS Enterprise Miner, with the highlighted nodes (figure 3.2.).

StatExplore

To get a good idea of the data, we took a quick glance at the variables' statistics through StatExplore.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	City_of_Residence	INPUT	8	0	Birmingham	14.72	Belfast	14.10
TRAIN	Department	INPUT	13	0	Psychiatry	13.60	General Practice	13.30
TRAIN	Education_Level	INPUT	9	29	Undergraduate	41.57	Master	33.88
TRAIN	Family_History	INPUT	5	0	Heart Disease	22.33	Hypertension	20.47
TRAIN	Gender	INPUT	3	0	Other	34.14	Female	33.77
TRAIN	Insurance_Provider	INPUT	6	104	Provider D	21.63	Provider A	20.03
TRAIN	Marital_Status	INPUT	4	0	Divorced	28.91	Single	28.56
TRAIN	Profession	INPUT	10	0	Retired	35.98	Student	20.96
TRAIN	Satisfaction_Level	INPUT	6	0	2	18.89	4	18.87

(*figure 3.3.*, StatExplore on categorical variables)

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	50.63565	31.18561	9952	56	0	52	195	0.28614	-0.21493
Approximate_Annual_Income	INPUT	43402.76	268142	9854	154	0	40874	11970900	42.21076	1835.703
Consultation_Duration	INPUT	67.80765	32.44714	10008	0	15	68	600	1.545225	21.11949
Consultation_Price	INPUT	187.263	862.8655	10008	0	50.03676	159.5248	39999.22	39.85869	1655.995
Insurance_Coverage	INPUT	115.4294	79.33776	9958	50	0	115.9291	421.8878	0.322906	-0.17236

(*figure 3.4.*, StatExplore on Numerical Variables)

Categorical Variables. In terms of category variability, all variables seem to not present any type of problem in terms of constancy or quasi-constancy. In other words, there are no variables with a single class.

In terms of missing values, we have two problematic variables: Education_Level and Insurance_Provider

- Education_Level is potentially due to lack in measurements, and it could be *"Missing at Random"*, as certain customers might have not been comfortable sharing such information.
- Insurance_Provider could be potentially due to non-applicability situations, meaning that some customers could have not had an insurance provider at all.

- There are six classes on Satisfaction_Level, when there should be five. This may suggest that a class which should not exist, is there (we will see later that it turns out to be level six)

We will consider imputing missing variables with a classifier, using predictive methods. This ensures that the imputation follows the existing patterns in the dataset, if they exist.

Numerical Variables. In the numerical variables we can already notice a few problems:

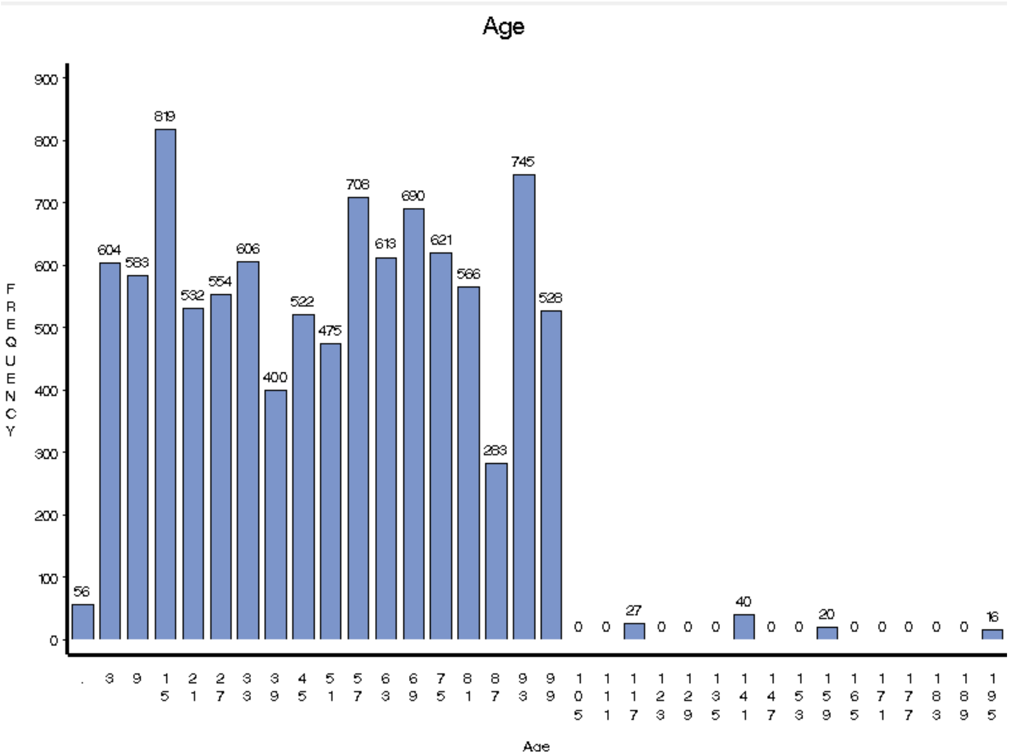
- In Age the maximum is 195, which is clearly an error in data measurement
- There are "extreme outliers" with Approximate_Annual_Income and Consultation_Price, as they have extremely high standard deviations: these could "ruin" our analysis of their distribution, which we will see in the next part.
- There are missing values in Age, Approximate_Annual_Income and Insurance_Coverage.

They will be imputed through a regressor, using predictive methods for the same reason described above.

MultiPlot

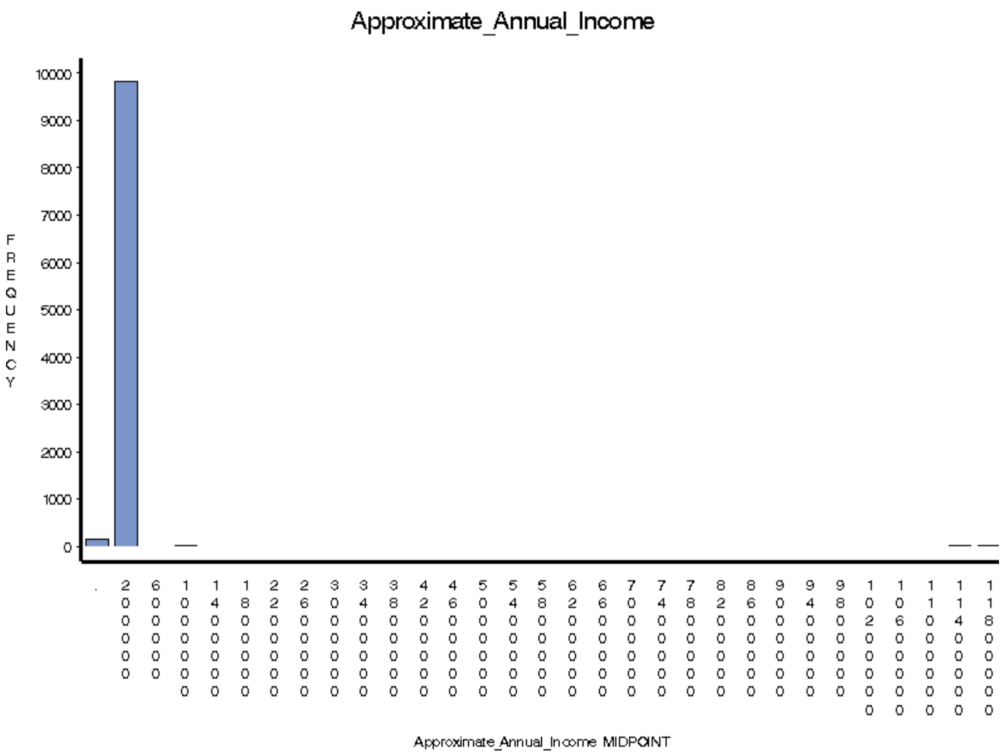
Successively, we looked at the variables' distributions through the MultiPlot node. Therefore, we will proceed on a case-by-case basis to analyze each variable.

- **Age**: As detected before, there are outliers with patients that have age > 111. Also, there are missing variables (56). The variable does not seem to follow any distribution, with some peaks favoring lower and higher ages. We will re-analyze this variable as we remove the outliers, to gain clearer insights.



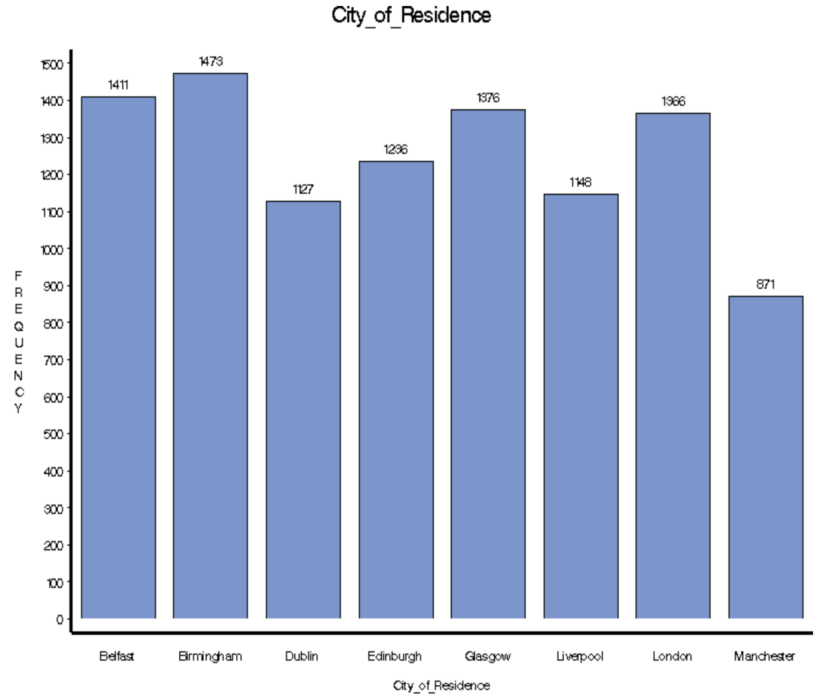
(*figure 3.5.*)

- **Approximate Annual Income**: In this case, the outliers are so "extreme" that it is impossible to analyze the variable's distribution; this is clearly an case of the "Bill Gates" effect. Also, as discussed previously, there are missing values.



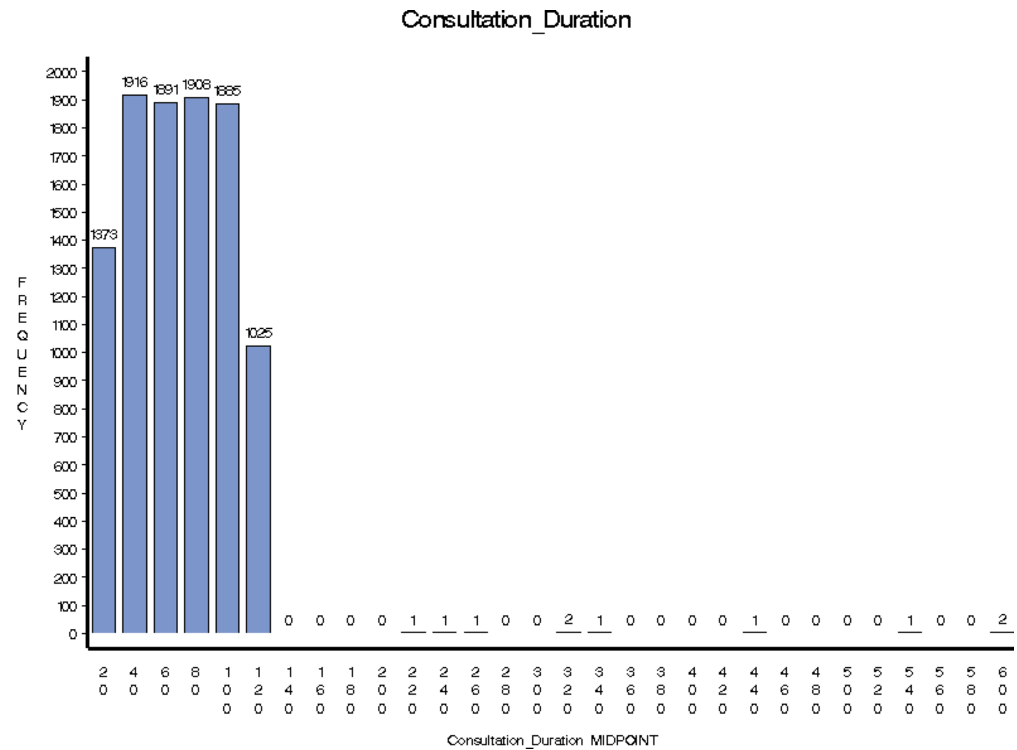
(*figure 3.6.*)

- **City of Residence**: No issues detected; cities of residence seem to be uniformly distributed between visitations.



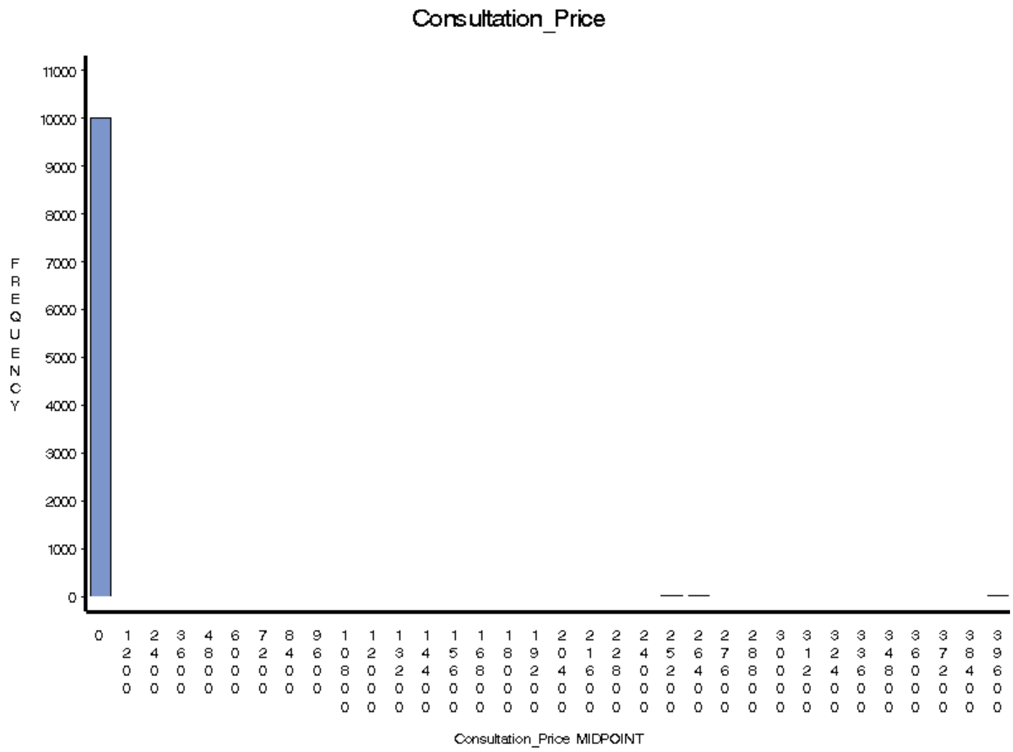
(*figure 3.7.*)

- **Consultation Duration**: Similarly to *Approximate Annual Income*, the outliers make it hard to analyze the variable's distribution: therefore, we will postpone the distribution's analysis to post-cleaning analysis. It might seem that this follows some sort of normal distribution. There are no missing values detected here.



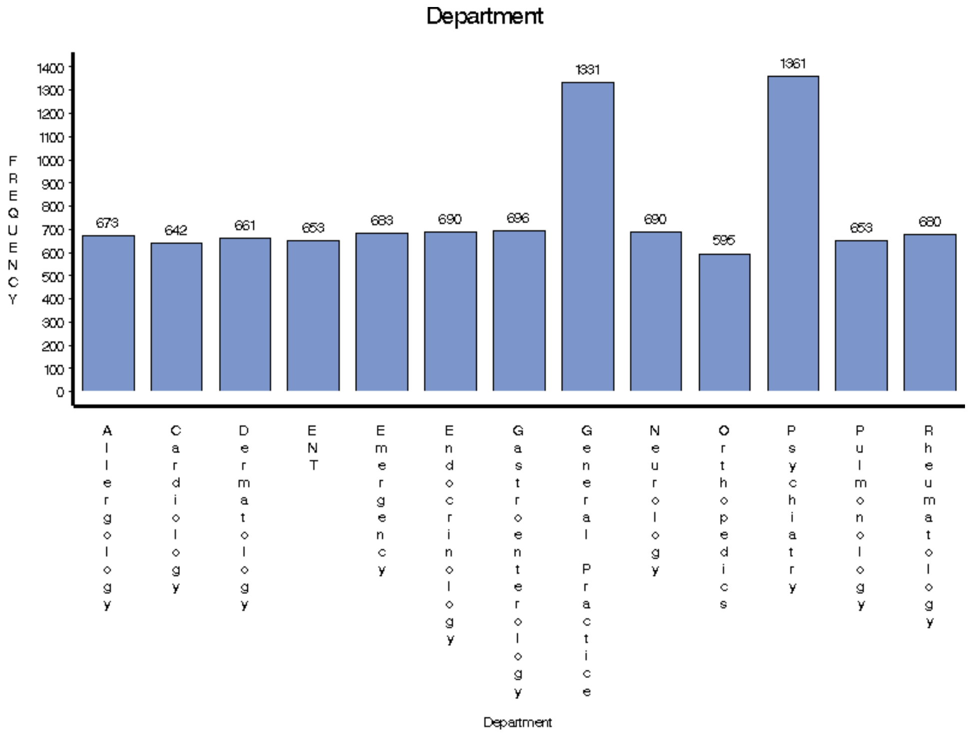
(*figure 3.8.*)

- **Consultation Price**: Same as above, the extreme outliers make it impossible to analyze the variable's distribution. So, we will postpone the analysis of this variable as nothing significant can be found.



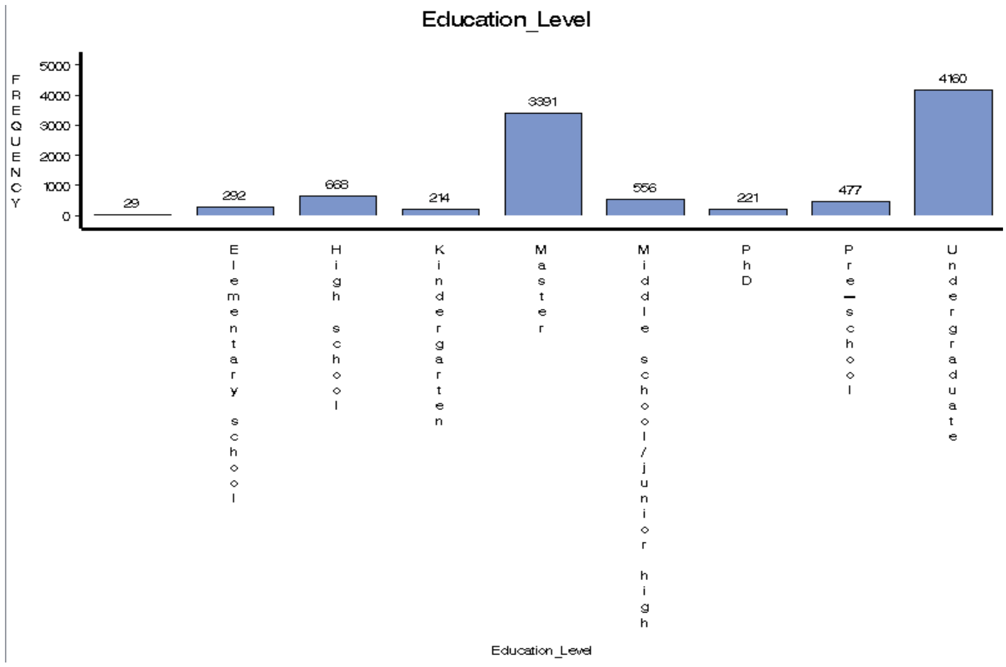
(*figure 3.9.*)

- **Department**: No issues detected. It seems to follow a uniform distribution, except for *General Practice* and *Psychiatry* departments as they have a slight peak, with more visits than everyone else. This clearly indicates that the two departments are the most popular ones for visits.



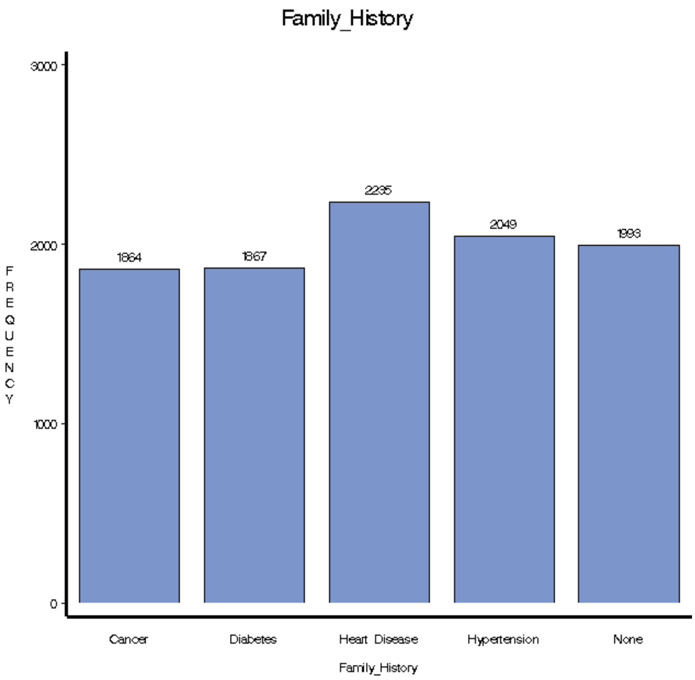
(*figure 3.10*)

- **Education Level**: Other than missing values (29), no problem is found. According to the distributions, it seems that there's a trend towards patients with master's or bachelor's degrees, occupying ~ 77.45% of the total transactions (or visits).



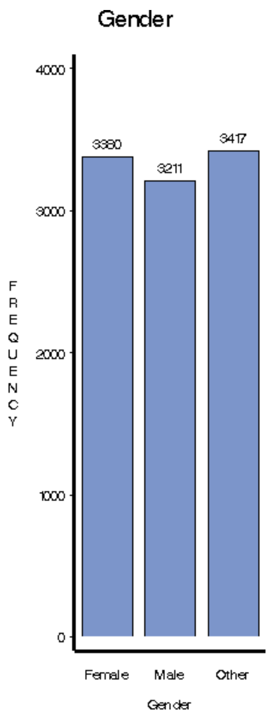
(*figure 3.11.*)

- **Family History**: The class *"None"* suggests that there might be missing values occupying a significant part of this variable (around 1/5ths); this might be a case of a missing variable being due to non-applicability, for instance cases of people whose family had no diseases. Therefore, if we consider *"None"* as a class of its own, we can say that this variable is uniformly distributed, with a slight trend towards heart disease.



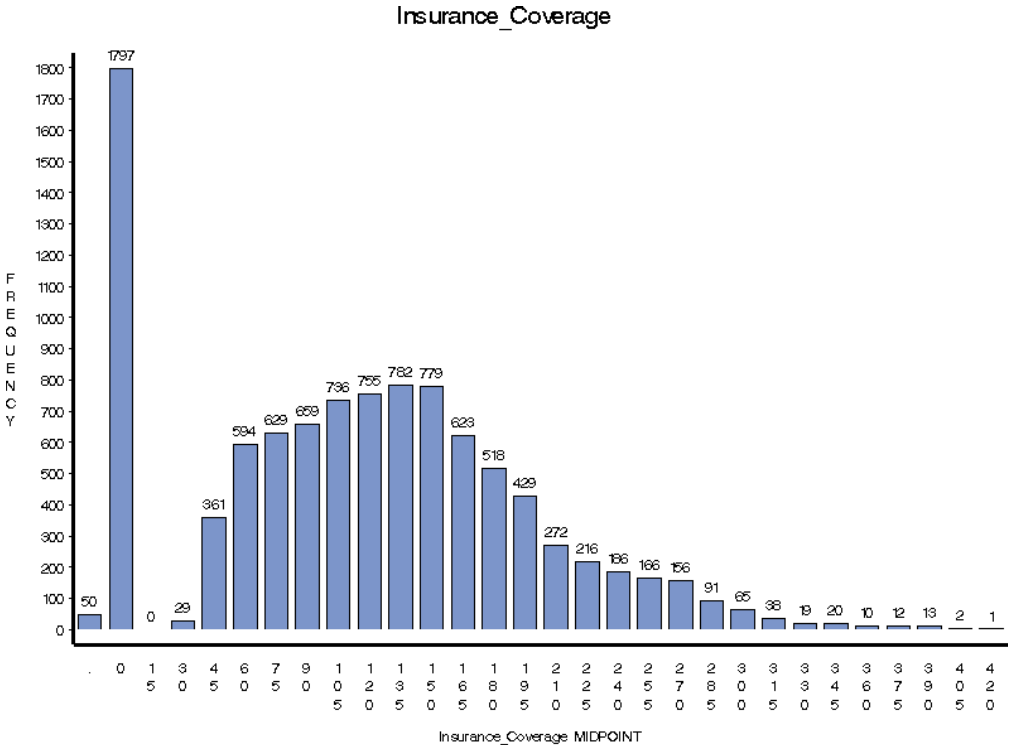
(*figure 3.12.*)

- **Gender**: No issues detected, variable follows a uniform distribution.



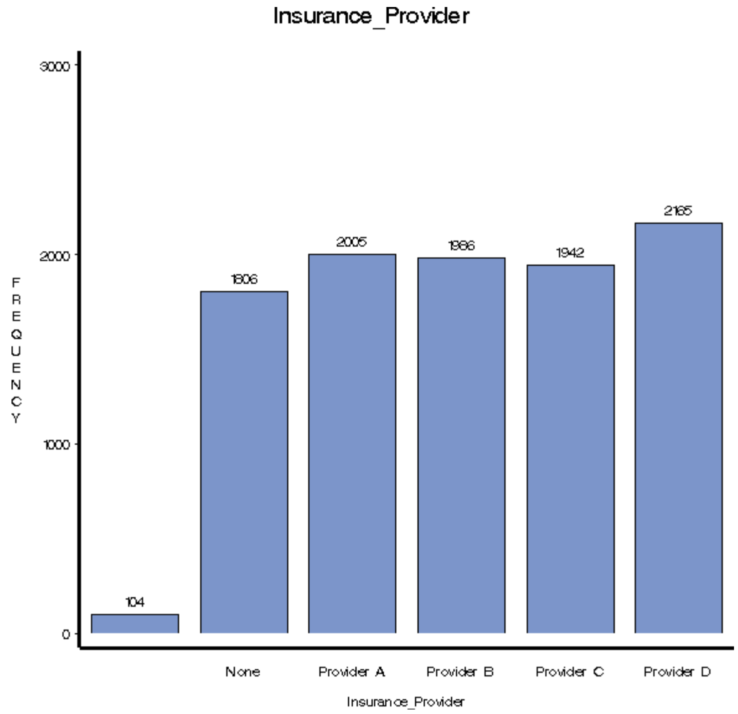
(*figure 3.13.*)

- **Insurance Coverage**: There are 50 missing values. We can gain an interesting insight about this variable: there's a peak of patients who had zero insurance coverage - potentially meaning that they had no insurance provider at all, as discussed previously - and ignoring this case, we have that the variable is slightly right-skewed.



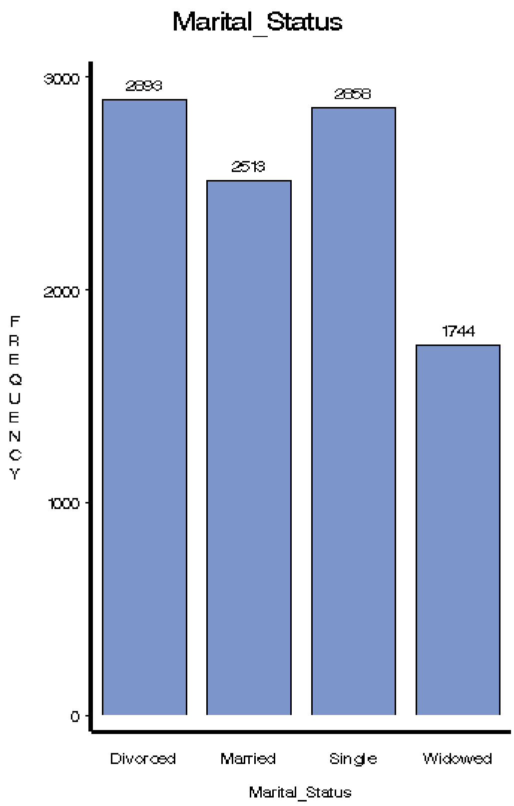
(*figure 3.14.*)

- **Insurance Provider**: There are missing values and the *"None"* class: meaning that missing values are not necessarily to be *"None"* class, as they could be caused by errors in data measurement. Other than that, insurance providers seem to be uniformly distributed.



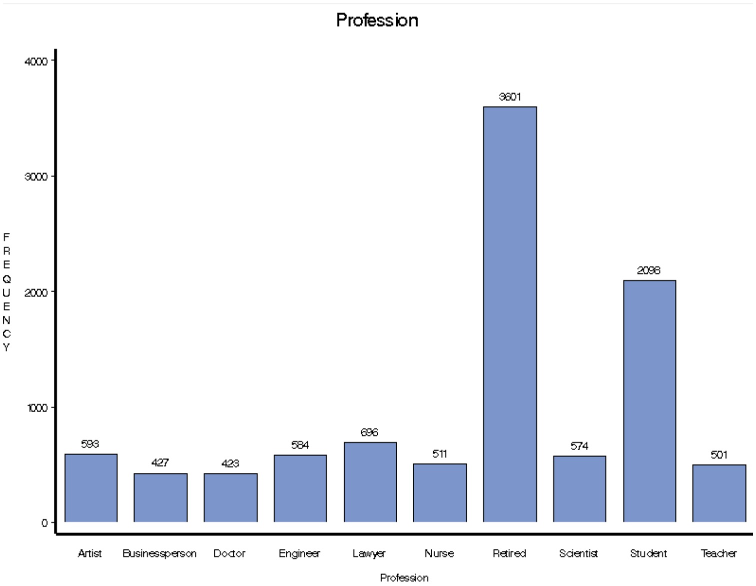
(*figure 3.15.*)

- **Marital Status**: No issues, there is a trend towards people who have been married (married, divorced and widowed). If we consider classes as their own, we cannot say anything about the classes' distribution.



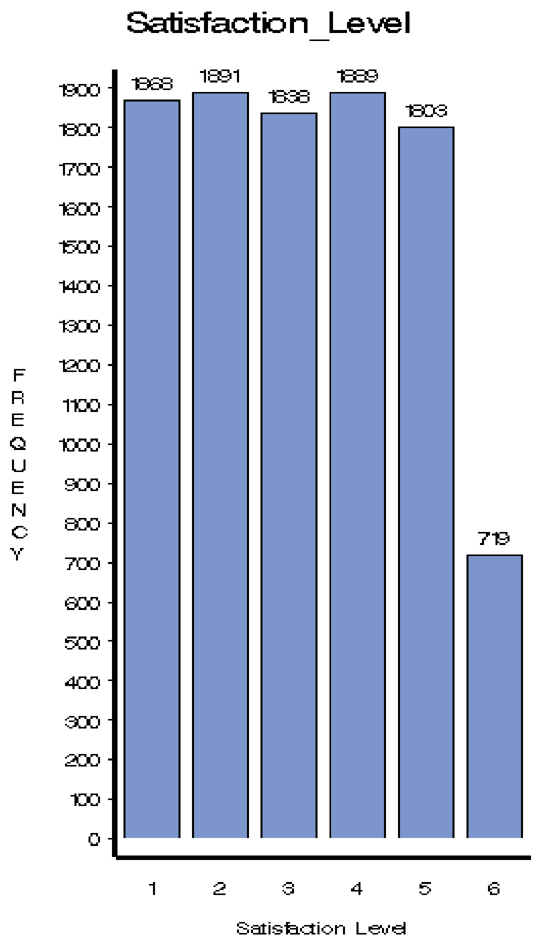
(*figure 3.16.*)

- **Profession**: No issues detected, classes other than "Retired" and "Student" seem to be uniformly distributed; there is a trend towards the two mentioned classes. This could suggest that most visits are either made by people of young or old age.



(*figure 3.17.*)

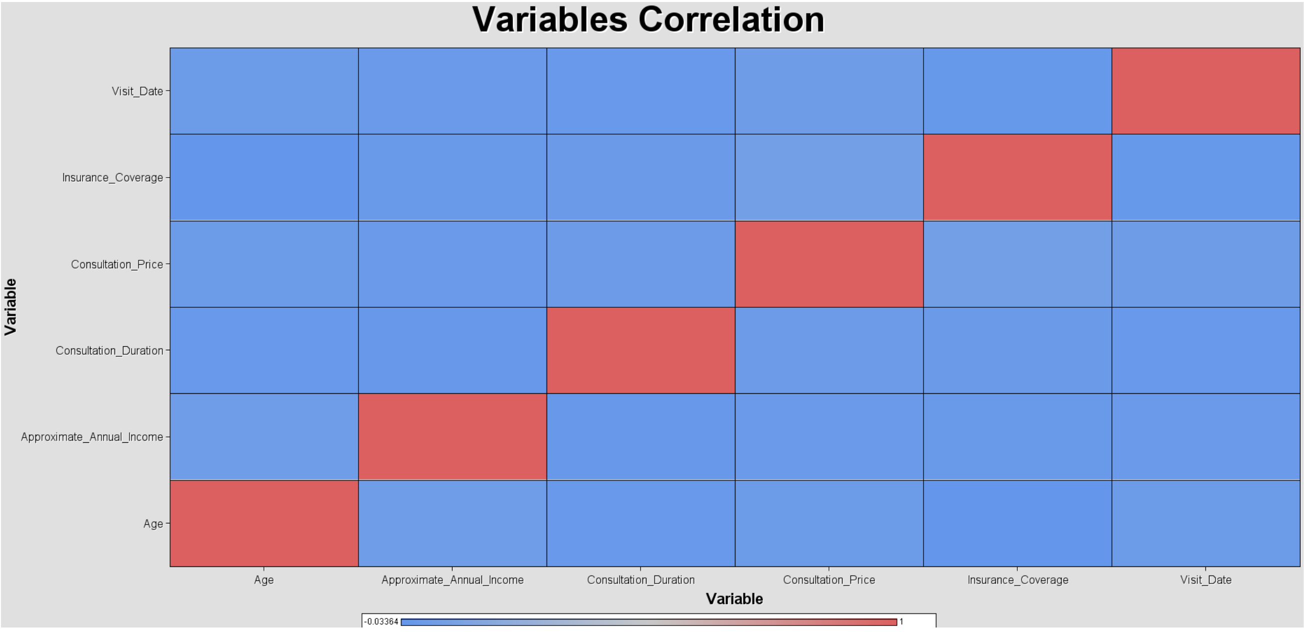
- **Satisfaction Level**: Classes seem to be uniformly distributed. However, there is a particular inconsistency between the existing classes and the metadata: the metadata suggests that levels should be from 1 to 5, meanwhile we there is class "6". We interpreted this class to be a non-answer, as customers could potentially choose not to communicate the satisfaction level of their visits.



(*figure 3.18*)

Variable Clustering

Lastly, we looked at the numerical variables' correlation with the *"Variable Clustering"* node. There seems to be no correlations, as all of them are inside the range $[-0.7, 0.7]$: all the correlation values seem to be near 0.033 (figure 3.19.), which indicates a low amount of correlation between numerical variables.
However, this result is to be re-checked as this result could have been "distorted" by the "dirtiness" of the data, namely outliers and missing values.



(*figure 3.19.*, Correlation Matrix for numerical variables)

Python

With Pandas' library in Python, we were able to extract information about the variable Visit_Date; it seems that all the visits happened in a time range from 1st January 2024 to 6th June 2024 (figure 3.20.). Therefore, we are talking about a time span of approximately 5 months; this insight will be relevant for data inconsistency checking purposes.

	Visit Date
count	10008
mean	2024-03-31 14:53:14.244604416
min	2024-01-01 00:00:00
25%	2024-02-15 00:00:00
50%	2024-03-31 00:00:00
75%	2024-05-16 00:00:00
max	2024-06-30 00:00:00

(*figure 3.20.*, Pandas' .describe() method on the Visit_Date variable)

Phase 1: Outliers and Missing Values Treatment

Let us remind the main problems with the data that have been detected in the previous phase:

- Outliers with Age
- Extreme outliers with Approximate Annual Income, Consultation Duration, Consultation Price
- Missing values with Age, Approximate Annual Income, Education Level, Insurance Coverage, Insurance Provider
- Unclear situation about Satisfaction Level, regarding inconsistent class.

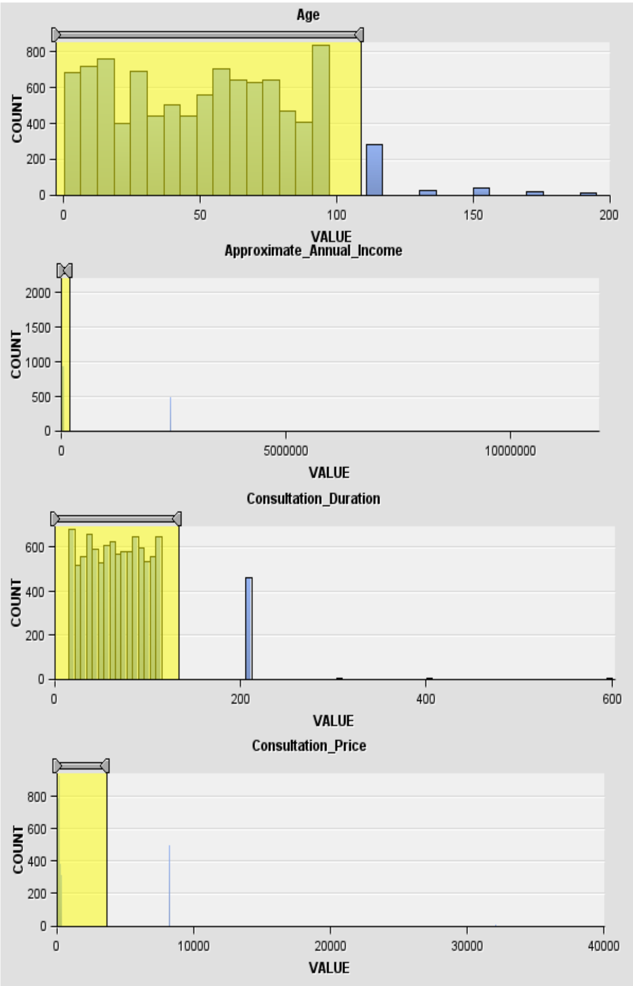
Outliers

(*figure 3.21.*, nodes used for outliers filtering)

Let us address the outliers first, to not cause any biased predictions during the imputation of missing values.

To deal with one-dimensional outliers, we manually defined a limit for each variable as a "filter range". In other words, we arbitrarily defined a range for which the variables would be classified as an outlier and thus be filtered from the main dataset. To do this, we used the *"Filter"* node (figure 3.21.).
In specifics, we have decided the following ranges (figure 3.22.):

- Age: $R \approx [0, 108]$
- Approximate Annual Income: $R \approx [0, 186740]$
- Consultation Duration: $R \approx [0, 133]$
- Consultation Price: $R \approx [0, 3636]$
-



(*figure 3.22.*, arbitrarily defined ranges)

As a result of this filtering, around 141 observations have been excluded from the dataset, which is approximately ~ 1.41% of the observations in the whole dataset.

We can consider this as a good number of observations to filter.

Number Of Observations

Data Role	Filtered	Excluded	DATA
TRAIN	9867	141	10008

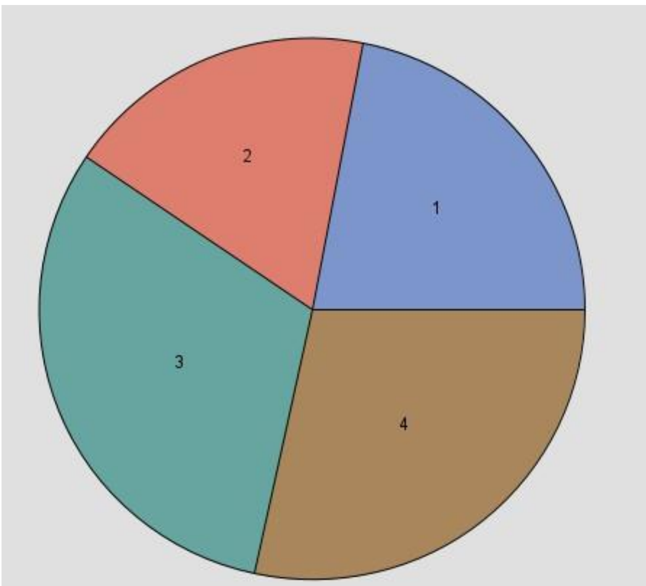
(*figure 3.23.*, summary of the filter)

Multidimensional Outliers

Before we impute values, we still need to check for multidimensional outliers. To do it, we used the *"Cluster Node"* (figure 3.21.) which performs K-means clustering on the dataset. This can be effective in finding these multidimensional outliers, as K-means is sensitive to them. More precisely, this node does the following:

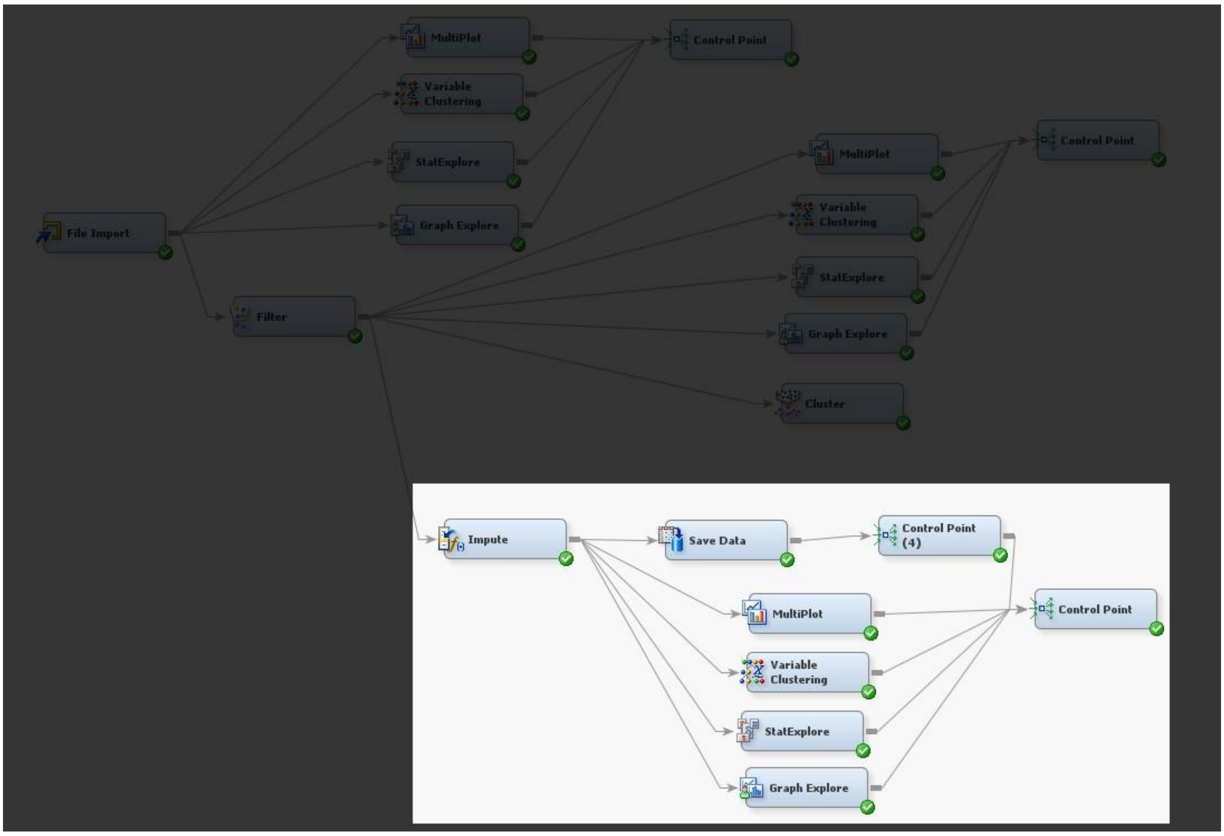
- Standardizes the numerical variables
- Initializes the seed with Princomp method, reducing the number of necessary iterations for the clustering process
- Makes four clusters: so, $K = 4$

  As a result, four quasi-equally sized clusters were formed, meaning there are no multidimensional outliers detected (figure 3.24.).



(*figure 3.24.*, result of 4-means clustering)

Missing Values



(*figure 3.25.*, nodes used for missing values imputation)

Having made sure that our data is clean from outliers, we can proceed to deal with missing values.

Then we decided to impute the missing values through *decision trees*, which can perform both classification and regression. We have not used KNN to perform imputation, as it is unavailable in the SAS Miner Enterprise program.

To perform this imputation, we used the *"Impute"* node (figure 3.25.), setting the method to "Decision Tree". It is worth noting that the imputed variables have been renamed to IMP_variable.

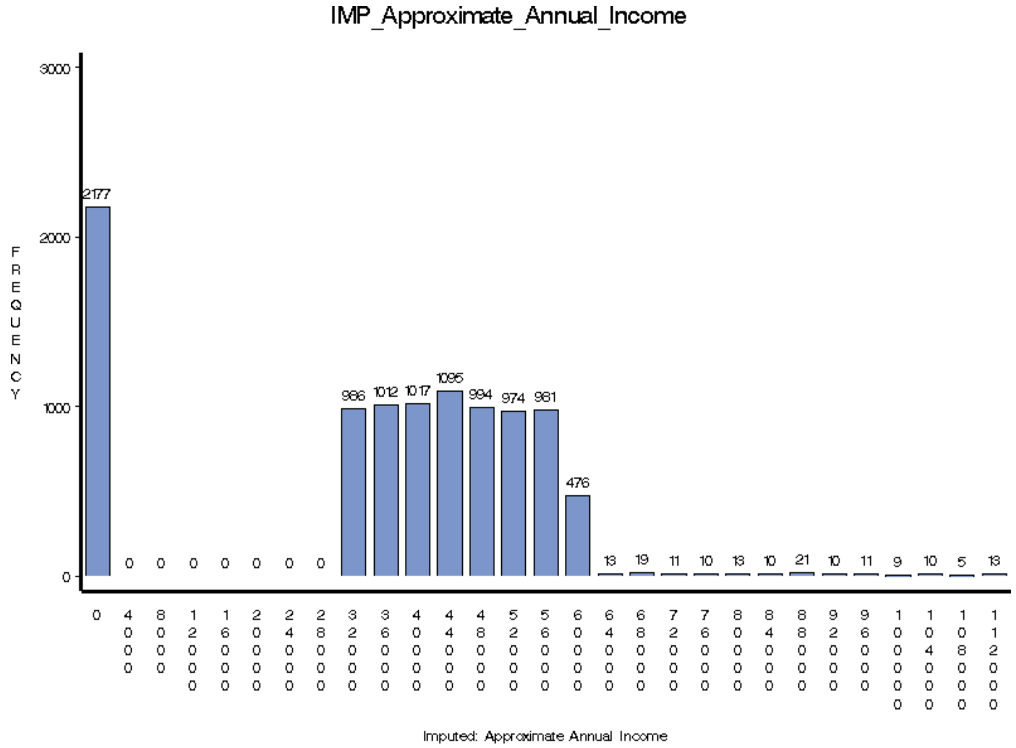
Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
Age	TREE	IMP_Age	.	INPUT	INTERVAL	Age	55
Approximate_Annual_Income	TREE	IMP_Approximate_Annual_Income	.	INPUT	INTERVAL	Approximate Annual Income	153
Education_Level	TREE	IMP_Education_Level	.	INPUT	NOMINAL	Education Level	29
Insurance_Coverage	TREE	IMP_Insurance_Coverage	.	INPUT	INTERVAL	Insurance Coverage	50
Insurance_Provider	TREE	IMP_Insurance_Provider	.	INPUT	NOMINAL	Insurance Provider	104

(*figure 3.26.*, results of tree imputation)

Post-Cleaning Analysis

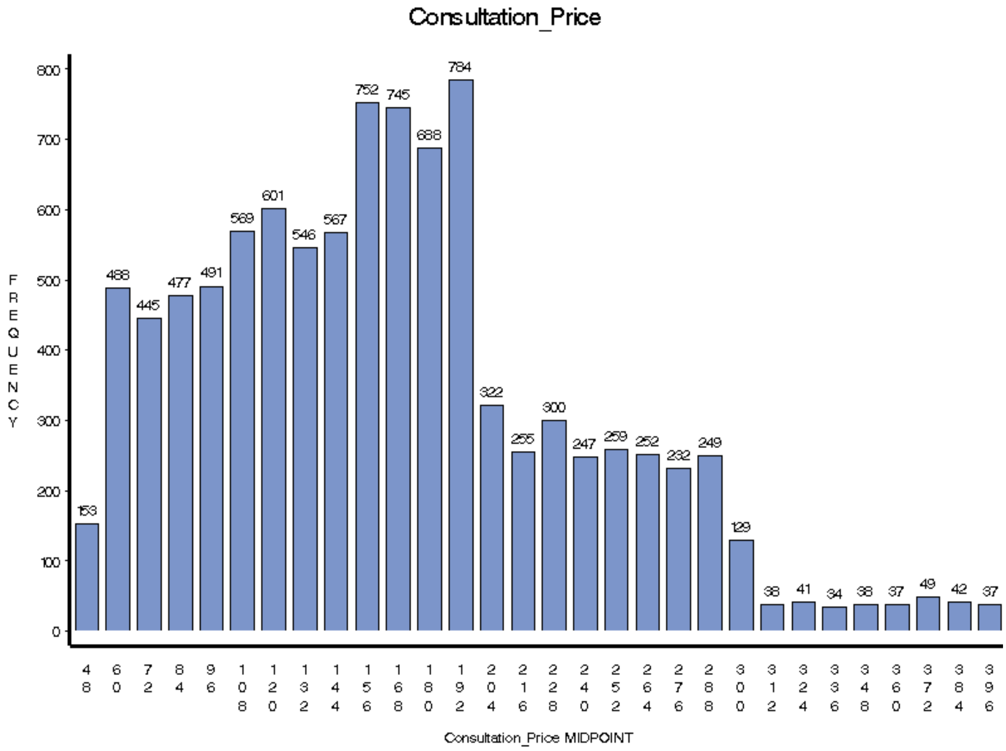
Having a clean dataset from outliers and missing values, we can check its statistics again. As remarked before, we will focus on the variables which were impossible to analyze due to extreme outliers - that is Approximate Annual Income, Consultation Duration and Consultation Price - and gain some immediate insights on the dataset.

- **Approximate Annual Income**: We can see an interesting fact: there is a neat separation between people with no income and people with income > 32.000. This could tell us that some of the patients were people who had no income at all, such as children or students. Other than that, the variable seems to be uniformly distributed, with some low-frequent values on the high range (they will not be considered as outliers as they are not "too far" from the values).



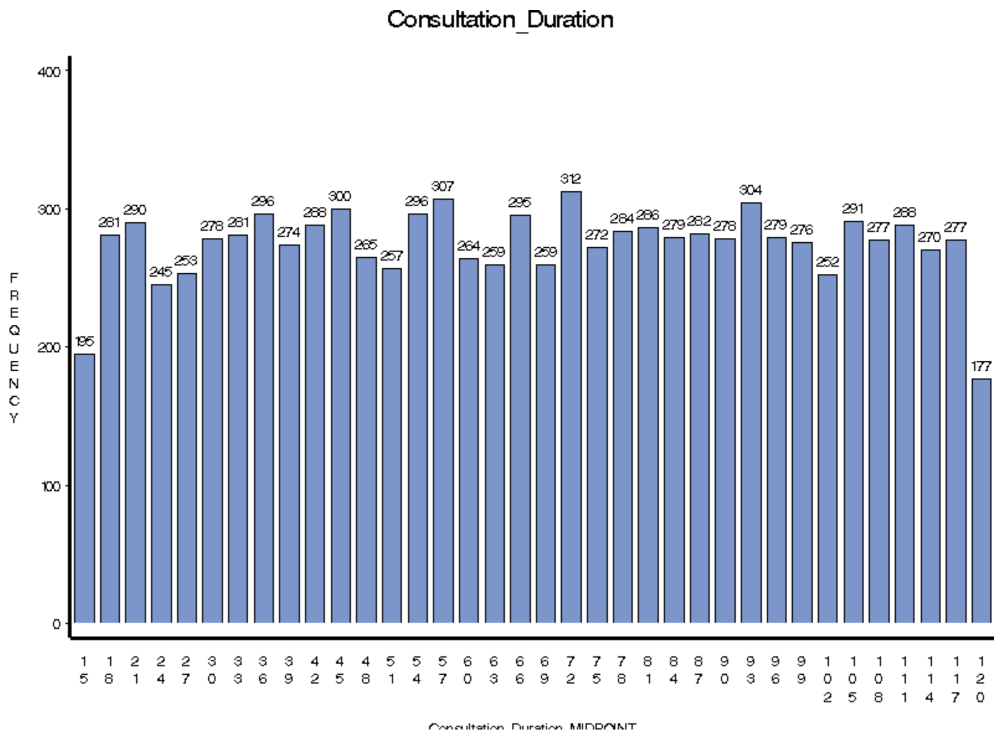
(*figure 3.27.*, cleaned)

- **Consultation Duration**: Without outliers, the consultation durations seem to be uniformly distributed, except for "extreme values" (first and last bin) which have a lower frequency.



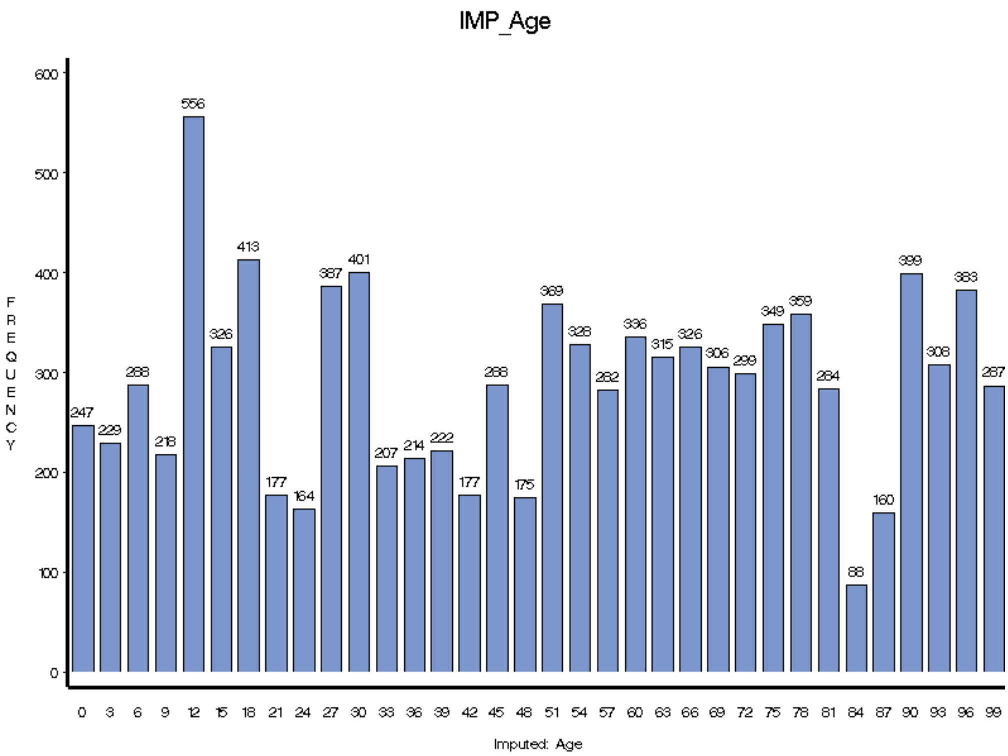
(*figure 3.28.*, cleaned)

- **Consultation Price**: The consultation prices seem to be distributed with a right-skew; this tells us that higher prices are rare (such as >312 pounds), whereas it's common to be charged around 150-200 pounds.



(*figure 3.29.*, cleaned)

- **Age**: Without outliers, we still cannot define a precise distribution for age; however, we can say that there is a trend towards people of young age (∈ [12, 15]); this confirms our previous hypothesis as we analyzed the approximate annual income, where most patients were underage people who cannot have an income.



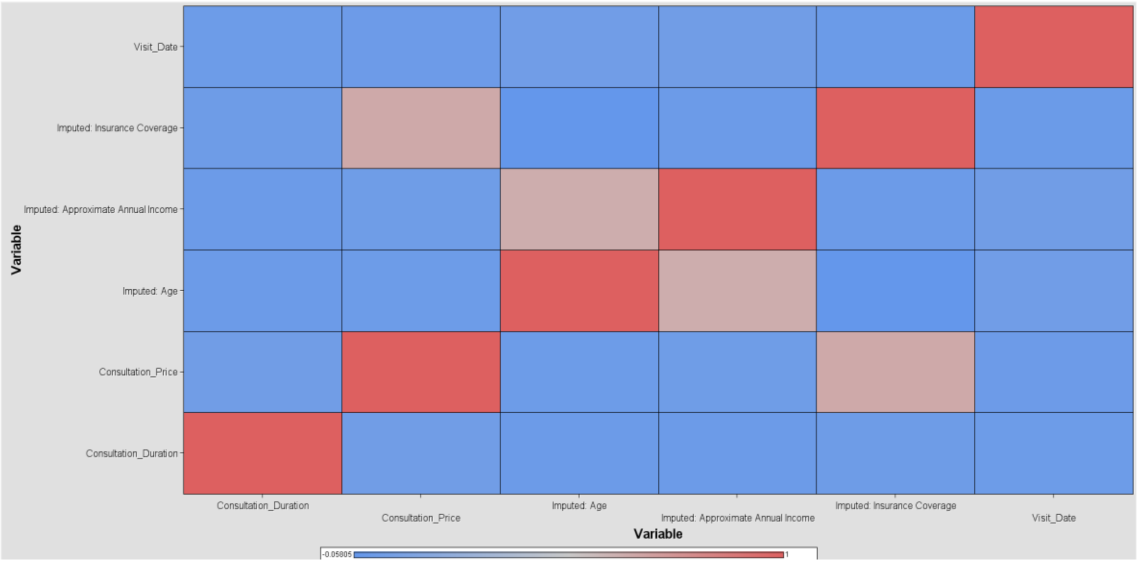
(*figure 3.30.*, cleaned)

Concerning the other variables, we can make the same conclusions as the ones we did previously (in *Phase 0*).

However, the situation becomes different if we check again the correlation between numerical variables. Here we obtain that there exist significant correlations. In fact, we can see that there is a significant amount of correlation between *Insurance Coverage* and *Consultation Price* (0.63), as well between *Approximate Annual Income* and *Age* (0.61) (figure 3.30.2). Although they're still inside the range [−0.7, 0.7], we still have potential grounds to consider these variables to be correlated enough. Both correlations make to make sense, as:

- Insurance providers usually cover a percentage of the consultation price, therefore if the consultation price is high then the insurance coverage is higher as well
- Age is correlated with annual income, as one's work career advances

As the project guidelines instructed, we will not do anything about the correlation and simply make it known in the report.



(*figure 3.30.2.*, correlation of variables post-data cleaning)

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	City_of_Residence	INPUT	8	0	Birmingham	14.46	Belfast	14.28
TRAIN	Department	INPUT	13	0	Psychiatry	13.65	General Practice	13.30
TRAIN	Family_History	INPUT	5	0	Heart Disease	22.33	Hypertension	20.25
TRAIN	Gender	INPUT	3	0	Female	33.95	Other	33.81
TRAIN	IMP_Education_Level	INPUT	8	0	Undergraduate	41.83	Master	34.08
TRAIN	IMP_Insurance_Provider	INPUT	5	0	Provider D	21.54	Provider B	20.53
TRAIN	IMP_REP_Satisfaction_Level	INPUT	5	0	4	22.72	2	19.96
TRAIN	Marital_Status	INPUT	4	0	Divorced	28.65	Single	28.54
TRAIN	Profession	INPUT	10	0	Retired	36.25	Student	20.96

(*figure 3.31.*, summary statistics of categorical variables)

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Consultation_Duration	INPUT	67.52133	30.47866	9867	0	15	68	120	-0.00403	-1.19091
Consultation_Price	INPUT	164.9322	71.24999	9867	0	50.03676	159.459	398.737	0.649235	0.182921
IMP_Age	INPUT	49.66622	29.67209	9867	0	0	52	100	0.00729	-1.23504
IMP_Approximate_Annual_Income	INPUT	35641.06	21094.31	9867	0	0	40979	113120	-0.56547	-0.20714
IMP_Insurance_Coverage	INPUT	115.0282	79.56521	9867	0	0	115.5196	421.8878	0.324597	-0.19115
Visit_Date	INPUT	2.0275E9	4544808	9867	0	2.0197E9	2.0275E9	2.0353E9	-0.01	-1.19551

(*figure 3.32.*, summary statistics of quantitative variables)

Final note: Variables Transformation

As specified in the guidelines, we will not standardize numerical variables. Moreover, we will not transform categorical variables to numerical with *one-hot encoding (or dummy transformation)*, as this could cause an excessive inflation in number of variables.

Phase 2: Data Inconsistencies Treatment

This phase of data treatment will make use of *SAS Guide* software, as this process may involve making some SQL queries.

Possible Inconsistencies

We thought out nine scenarios of data inconsistency, and they are:

i. Age should be above 0
ii. Legal marriage in the United Kingdom is 18; so, anyone under 18 who presents marital status other than single is considered as an anomaly
iii. School leaving age is defined to be 16 in the United Kingdom; therefore anyone ≤ 16-aged customers should be a student
iv. Insurance coverage should be always smaller (or equal) than the consultation price
v. People without an insurance provider should not have insurance coverage at all
vi. Some professions might require some degrees; in our case, we considered Engineers, Lawyers and Scientists to be at least undergraduates (or higher).
vii. Students should not possess an income
viii. Ages and education level should coincide; in particular, some education levels have an intrinsic "minimum age". We considered them as the following:

- You need to be at least 16 to have a high school
- diploma You need to be at least 21 to have a
  bachelor's degree
- You need to be at least 22 to have a master's degree; in United Kingdom master's degrees last one year
- You need to be at least 25 to have a PhD

  In this case, we allowed a discrepancy of one year to account for people who started school one year earlier or later.

ix. People with paying jobs (e.g. not Student nor Retired) should have an income

In any case of inconsistency, rows will be deleted.

Some variables should remain constant between patients (Profession, Age, Gender,

Family History, Insurance Provider, Marital Status and City of Residence) should remain the same. To do this, we will use SQL queries and proceed on a case-by-case basis.
The reason we are checking this, is that the timespan of the dataset is around five months (figure 3.20), and the previously mentioned variables should not vary in such a short time span.
As an end-result, this makes possible to build ABTs without any type of inconsistencies.

Results
The code to treat the first nine scenarios of data inconsistency was written in SAS code, and we filtered out inconsistent data in the following order:

1. Age
2. Satisfaction Level
3. Age and Marriage
4. Age and Profession=Student
5. Satisfaction Value
6. Age and Marital Status
7. Insurance Coverage and Consultation Cost
8. Insurance Provider and Insurance Coverage
9. Education Level and Profession
10. Profession=Student and Approximate Annual Income
11. Age and Education Level
12. Profession and Annual Income

As a result, we have the following sequence which represents the decrease in instance as we check for inconsistent rows:

$$9867 \xrightarrow{1.} 9724 \xrightarrow{2.} 9724 \xrightarrow{3.} 9719 \xrightarrow{4.} 9610 \xrightarrow{5.} 9599 \xrightarrow{6.} 9599$$
$$9599 \xrightarrow{7.} 9599 \xrightarrow{8.} 9599 \xrightarrow{9.} 9599 \xrightarrow{10.} 9599 \xrightarrow{11.} 9389 \xrightarrow{12.} 9389$$

Therefore, from these series of controls we have deleted 478 rows, which is the ~ 4.85% of the original dataset size.

Concerning the controls about the "constant" variables between patients' IDs, we have the following result:

- **Profession:** Two patients had inconsistencies in profession: they are the ones with ID 1488 and 1496. Looking at their age, their profession should be corrected to "Student" (figure 3.33.); it might be that there were visits where his profession was erroneously classified as "Retired". We will manually correct them to be defined as "Student" in another SAS script.

- **Age**: There were a lot of inconsistencies in age, mainly due to tree-imputation. As the values are "close to each other", we can consider doing nothing about them and taking the mean for building the ABT.
- **Gender**: There were five patients with inconsistent genders: 1050, 1307, 1349, 1447, 1490. The fact the difference in genders do not follow a consistent timeline suggests that this is due to a registration error, rather than gender transitioning (figure 3.34).
  Therefore, their genders will be replaced by the mode of each patient's gender.
- **Family History**: No inconsistencies detected
- **City of Residence**: No inconsistencies detected
- **Marital Status**: The following patients had inconsistent marital status: 1140, 1322, 1332, 1382. By analyzing their marital statuses row-by-row, we have found out that each patient with inconsistent marital status had only one row with mismatching information (figure 3.35.).
  Therefore, we ruled this to be due to registration error, rather than transitioning; so, the inconsistencies will be replaced with the correct value.
- **Insurance Provider**: Interestingly enough, there are a good number of patients with different insurance providers for each visit. This makes sense, as certain patients - such as children - can benefit from multiple insurance providers. There are 31 patients with different insurance providers, and they make up 612 rows of the dataset (so around 6.20% of the total transactional dataset). It is possible to separate them into another date for special analysis, as they make up a significant amount of data. For our ABT, we will filter these rows out but the transactional rows will be kept in the original transactional table, as they are not inconsistent in the sense that they are due to errors.

All of this has been done with SAS code (see appendix.)

Patient_ID	Profession	Profession	Visit_Date	IMP_Age
1488	Retired	Student	15APR2024:00:00:00	11
1488	Retired	Student	08MAY2024:00:00:00	11
1488	Student	Retired	02JAN2024:00:00:00	11
1488	Student	Retired	02FEB2024:00:00:00	11
1488	Student	Retired	11FEB2024:00:00:00	11
1488	Student	Retired	23FEB2024:00:00:00	11
1488	Student	Retired	26FEB2024:00:00:00	11
1488	Student	Retired	13MAR2024:00:00:00	11
1488	Student	Retired	15MAR2024:00:00:00	11
1488	Student	Retired	19MAR2024:00:00:00	11
1488	Student	Retired	09APR2024:00:00:00	11
1488	Student	Retired	12APR2024:00:00:00	11
1488	Student	Retired	06MAY2024:00:00:00	11
1488	Student	Retired	09MAY2024:00:00:00	11
1488	Student	Retired	10MAY2024:00:00:00	11
1488	Student	Retired	17MAY2024:00:00:00	11
1488	Student	Retired	30MAY2024:00:00:00	11
1488	Student	Retired	10JUN2024:00:00:00	11
1496	Retired	Student	05JAN2024:00:00:00	4
1496	Retired	Student	25JAN2024:00:00:00	4
1496	Retired	Student	21FEB2024:00:00:00	4
1496	Retired	Student	11MAR2024:00:00:00	4
1496	Retired	Student	22MAR2024:00:00:00	4
1496	Retired	Student	28MAR2024:00:00:00	4
1496	Retired	Student	15APR2024:00:00:00	4
1496	Retired	Student	22APR2024:00:00:00	4

(*figure 3.33*, customers with inconsistent profession)

Patient_ID	Gender	Visit_Date
1050	Female	08JAN2024:00:00:00
1050	Female	15JAN2024:00:00:00
1050	Female	13FEB2024:00:00:00
1050	Female	23FEB2024:00:00:00
1050	Female	29FEB2024:00:00:00
1050	Female	09MAR2024:00:00:00
1050	Female	27MAR2024:00:00:00
1050	Female	29MAR2024:00:00:00
1050	Female	23APR2024:00:00:00
1050	Female	03MAY2024:00:00:00
1050	Female	10JUN2024:00:00:00
1050	Female	13JUN2024:00:00:00
1050	Female	14JUN2024:00:00:00
1050	Female	15JUN2024:00:00:00
1050	Male	15JAN2024:00:00:00
1050	Male	28JAN2024:00:00:00
1050	Male	13MAR2024:00:00:00

(*figure 3.34.*, example of a customer with inconsistent gender)

Patient_ID	Marital_Status	Marital_Status	Visit_Date
1140	Single	Widowed	10JAN2024:00:00:00
1140	Single	Widowed	29JAN2024:00:00:00
1140	Single	Widowed	30JAN2024:00:00:00
1140	Single	Widowed	04FEB2024:00:00:00
1140	Single	Widowed	10FEB2024:00:00:00
1140	Single	Widowed	19FEB2024:00:00:00
1140	Single	Widowed	18MAR2024:00:00:00
1140	Single	Widowed	03APR2024:00:00:00
1140	Single	Widowed	06APR2024:00:00:00
1140	Single	Widowed	11APR2024:00:00:00
1140	Single	Widowed	17APR2024:00:00:00
1140	Single	Widowed	29APR2024:00:00:00
1140	Single	Widowed	20MAY2024:00:00:00
1140	Single	Widowed	07JUN2024:00:00:00
1140	Single	Widowed	18JUN2024:00:00:00
1140	Single	Widowed	29JUN2024:00:00:00
1140	Widowed	Single	22FEB2024:00:00:00
1322	Married	Single	25JAN2024:00:00:00
1322	Married	Single	23JUN2024:00:00:00
1322	Single	Married	05JAN2024:00:00:00
1322	Single	Married	12JAN2024:00:00:00
1322	Single	Married	13JAN2024:00:00:00
1322	Single	Married	19JAN2024:00:00:00
1322	Single	Married	22JAN2024:00:00:00
1322	Single	Married	19FEB2024:00:00:00
1322	Single	Married	28FEB2024:00:00:00

(*figure 3.33*, example of a customer with inconsistent marital status)

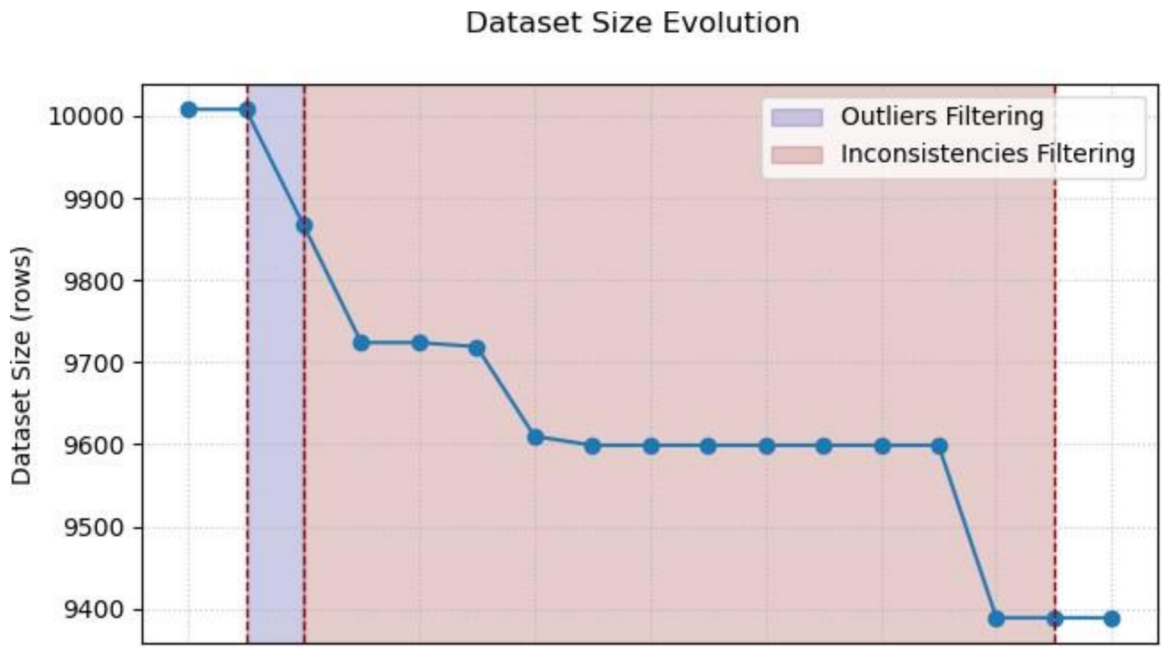
Results

As a result, we can consider the whole dataset to be ready for analysis as it is clean from any kind of "dirtiness", e.g. outliers, missing values, inconsistent rows. Looking at the whole process in its retrospect, the dataset went through a shrinkage of entries, from 10008 to 9389 rows.
With the preprocessing pipeline we removed the ~ 6.185% of the original dataset (619 rows).

Final Touches with Excel

With Excel, we made some final touches to the final transactional table, which will be used in the *"Data Visualization"* part.

- Renamed every column of type IMP_var or IMP_REP_var to var
- Removed *WARN* column
- Made new column named SATISFACTION_LEVEL_NEW, where class six is replaced with N/A, for data visualization purposes; in this way, it's clearer that satisfaction level 6 represents a non-answer, rather than any other time of anomaly. The formula used is =IF(cell=6, "N/A", cell)



(*figure 3.34.*, evolution of dataset size)

ABT Construction

The modified dataset obtained remains a *transactional table*, meaning we still have no insights about the *customers itself*. To obtain a source of data where we can glean insights about customers, we'll have to transform the transactional table into an analytic-base table.

Before proceeding to construct the ABT, we will deal with satisfaction level class six, as it is some sort of non-answer; therefore, before building the ABT we will delete every row with satisfaction level six, in order not to "distort" the derived variables for the ABT.

To do this, we will derive the following variables from the transactional table:

Pivoting: We can directly transpose some variables to each customer, which we assumed to be unique. They are namely gender, profession, marital status, city of residence, family history and insurance provider.

Aggregation: We can get frequency, recency, membership and monetary of the customer.

- *Frequency* is the total amount of transactions linked to a patient
- *Recency* is the amount of days since the last visit
- *Membership* is the amount of days since the first visit
- *Monetary* is the total sum of consultation price

Summarization: We can get the following averages:

- *Average Approximate Annual Income*
- *Average Age*: there were some mismatches in age, due to imputations. As previously established, we can do this as the values are "near" enough.
- *Average Satisfaction*
- *Level Average*
  Consultation Duration

Proportions. We can get the following proportion:

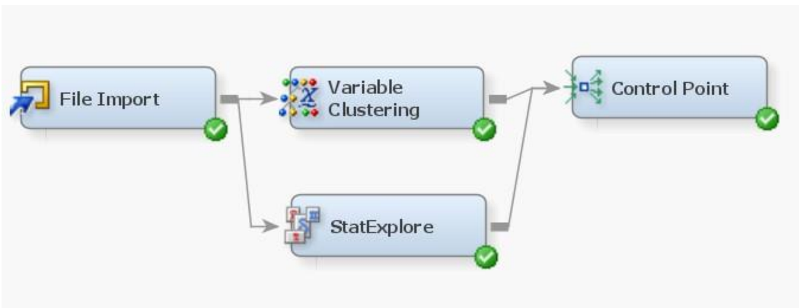
- *Total insurance coverage* respect to *total charged amount* for all visits of a patient

Segmentation of Departments: We can segment each department visit to get the following information:

- Amount of consultations done, relative to the frequency of a patient
- Proportion of prices, relative to monetary of a patient

Statistical Analysis

Before proceeding to do a complete data visualization of the ABT's data, we will make some quick remarks on the ABT's statistical properties. To do this, we used SAS Enterprise Miner to obtain final statistics and the correlation matrix.



(*figure 4.0.*, SAS Enterprise Miner workflow)

Summary Statistics: Regarding the categorical variables, we can say that most demographics is composed by retired people and students, as they make up ≈ 56.28% of the customer dataset (*figure 4.1.*).
Looking at the summary statistics of the numerical variables, we can obtain a lot of significant insights. For example, the patients seem to be averagely satisfied with the hospital, with an average rating of 3.124 with a low variability (standard deviation of around 0.384, therefore it is expected for the ratings to fall in the range [1.972, 4.276] according to the 3-σ rule).
Also, we can say that there are patients whose consultation charges covered only the *Emergency* department; this means that there are patients who went to the hospital only for emergency-related issues.

Correlation Matrix: The first thing we can notice is that there is a "diagonal" of strongly correlated variables (∀ ~ 0.9) (figure 4.3.). These are the variables which are derived by

segmenting departments into the proportion of total monetary and frequencies (figure 4.4.). This makes sense, as having more than others means that they get paid the most by the customer. Another correlation we noticed is that there's a strong correlation between monetary and frequency (~ 0.9) (figure 4.4.); this also makes sense, as more visits imply paying more and more for each visit.
The last correlation, which is also the weakest, is the one between age and average recorded annual income (~ 0.6) (figure 4.3.).

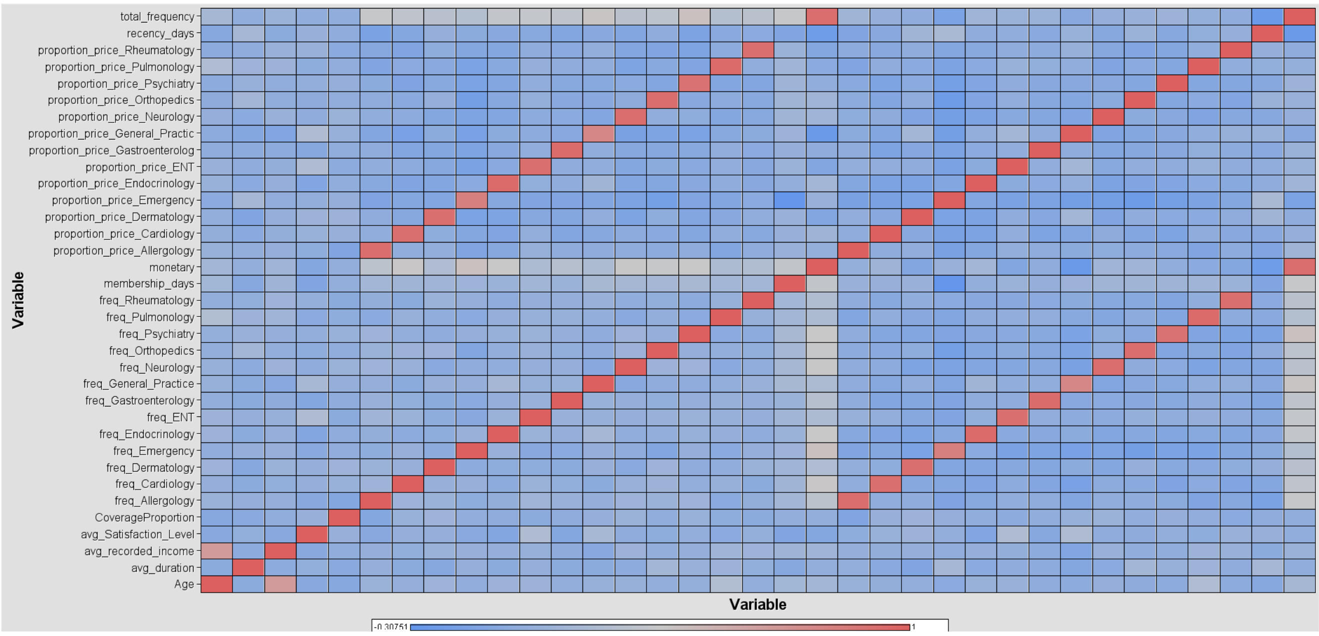
These aspects will be widely explored in the *"Data Visualization"* phase; in other words, this part served the project as a sort of outline for the next part.

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	City_of_Residence	INPUT	8	0	Birmingham	14.57	Belfast	14.35
TRAIN	Family_History	INPUT	5	0	Heart Disease	22.65	Diabetes	20.40
TRAIN	Gender	INPUT	3	0	Other	34.75	Female	33.18
TRAIN	IMP_Insurance_Provider	INPUT	5	0	Provider B	21.30	Provider C	21.08
TRAIN	Marital_Status	INPUT	4	0	Divorced	29.60	Single	27.35
TRAIN	Profession	INPUT	10	0	Retired	36.10	Student	20.18

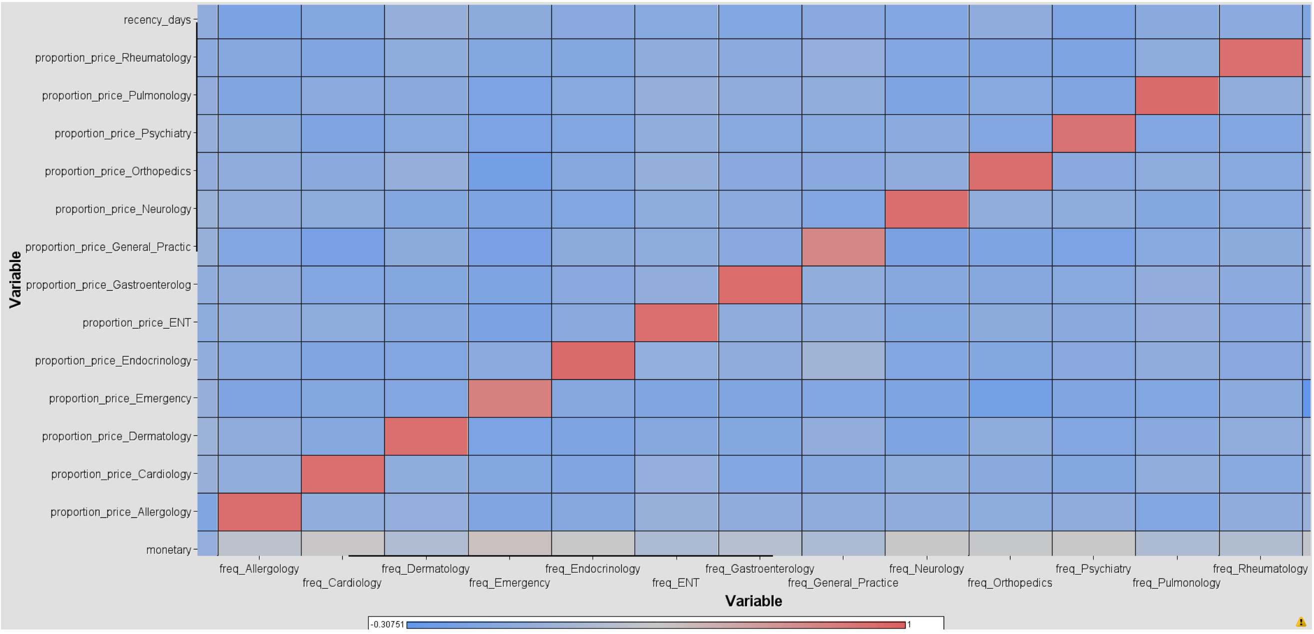
(*figure 4.1.*, summary statistics of categorical variables)

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Age	INPUT	50.36791	28.84006	446	0	1	52	100	0.009775	-1.20008
CoverageProportion	INPUT	0.682243	0.353396	446	0	0	0.8	1	-1.22671	-0.07077
avg_Satisfaction_Level	INPUT	3.214816	0.384139	446	0	2.227273	3.190476	5	0.254172	0.90518
avg_duration	INPUT	67.50553	6.922461	446	0	48.36842	67.61905	96	0.084845	0.381289
avg_recorded_income	INPUT	36510.52	19159.89	446	0	0	44171.99	88209.81	-1.07845	0.30301
freq_Allergology	INPUT	1.318386	1.07556	446	0	0	1	5	0.656565	0.01843
freq_Cardiology	INPUT	1.280269	1.119758	446	0	0	1	7	0.935776	1.341578
freq_Dermatology	INPUT	1.26009	1.151286	446	0	0	1	5	0.819398	0.13045
freq_ENT	INPUT	1.26009	1.123627	446	0	0	1	5	0.763418	0.092969
freq_Emergency	INPUT	1.331839	1.263923	446	0	0	1	7	0.945806	0.804936
freq_Endocrinology	INPUT	1.367713	1.212014	446	0	0	1	6	0.93911	0.70147
freq_Gastroenterology	INPUT	1.367713	1.230415	446	0	0	1	6	0.805429	0.229019
freq_General_Practice	INPUT	2.591928	1.60598	446	0	0	2	9	0.565338	0.136675
freq_Neurology	INPUT	1.363229	1.158415	446	0	0	1	6	0.897277	0.771365
freq_Orthopedics	INPUT	1.161435	1.153769	446	0	0	1	6	1.1188	1.33893
freq_Psychiatry	INPUT	2.717489	1.734257	446	0	0	3	11	0.754913	1.26453
freq_Pulmonology	INPUT	1.298206	1.144934	446	0	0	1	6	0.850228	0.486783
freq_Rheumatology	INPUT	1.318386	1.116565	446	0	0	1	5	0.797325	0.572552
membership_days	INPUT	331.9592	12.22163	446	0	175.5982	334.5982	340.5982	-6.33602	66.37483
monetary	INPUT	3241.079	834.0827	446	0	266.8861	3225.102	6189.114	0.124453	0.801259
proportion_price_Allergology	INPUT	0.060381	0.051824	446	0	0	0.051055	0.297938	0.840587	0.692987
proportion_price_Cardiology	INPUT	0.08872	0.076163	446	0	0	0.080402	0.393042	0.824632	0.758999
proportion_price_Dermatology	INPUT	0.059129	0.05609	446	0	0	0.048873	0.277781	1.020274	0.947046
proportion_price_ENT	INPUT	0.029875	0.027856	446	0	0	0.025852	0.151965	1.034197	1.244222
proportion_price_Emergency	INPUT	0.121515	0.122286	446	0	0	0.105549	1	2.029601	10.64956
proportion_price_Endocrinology	INPUT	0.092493	0.077614	446	0	0	0.081714	0.36141	0.641311	-0.03323
proportion_price_Gastroenterolog	INPUT	0.063506	0.059759	446	0	0	0.055317	0.349561	1.190717	2.075078
proportion_price_General_Practic	INPUT	0.063243	0.047923	446	0	0	0.056013	0.581491	3.330204	30.13581
proportion_price_Neurology	INPUT	0.094078	0.078843	446	0	0	0.081414	0.396555	0.775637	0.382476
proportion_price_Orthopedics	INPUT	0.078632	0.078333	446	0	0	0.067824	0.470979	1.210594	2.282459
proportion_price_Psychiatry	INPUT	0.126493	0.079383	446	0	0	0.122593	0.518467	0.749309	1.440611
proportion_price_Pulmonology	INPUT	0.060706	0.055133	446	0	0	0.05158	0.267471	0.890992	0.423555
proportion_price_Rheumatology	INPUT	0.06123	0.051833	446	0	0	0.053866	0.251601	0.718629	0.192402
recency_days	INPUT	168.3135	9.938331	446	0	159.5982	164.5982	245.5982	2.464708	10.03215
total_frequency	INPUT	19.63677	4.700743	446	0	1	20	39	0.001487	1.068595

(*figure 4.2.*, summary statistics of quantitative variables)



(*figure 4.3.*, matrix correlation of ABT's variables)

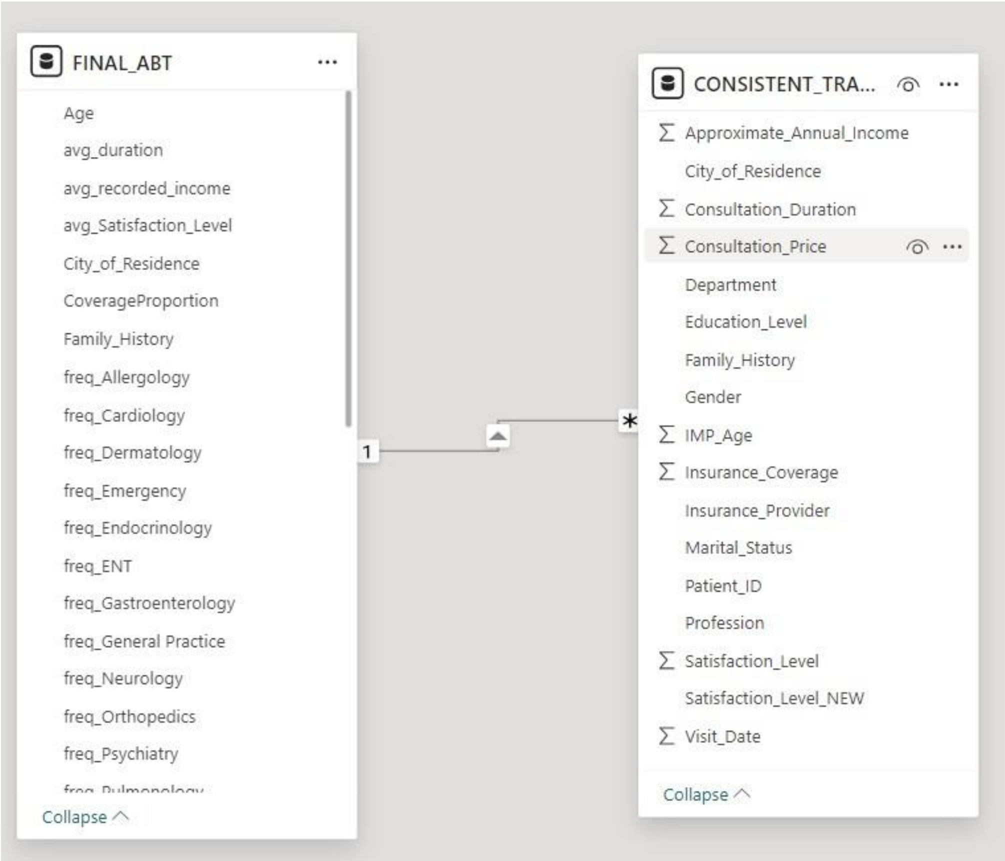


(*figure 4.4.*, zoom on *fig. 4.1.*)

With this, we can finally conclude the *"Modify"* process of the SEMMA pipeline.

Data Visualization with PowerBI

To gain as many insights as possible, we have decided to make the most of our data available. That is, we used the final ABT and the preprocessed transactional table; we integrated them with each other as they are related through a *"One-to-many"* relationship, with each entry in the signature table having one or many entries in the transactional table. In other words, a patient has one or more visits.

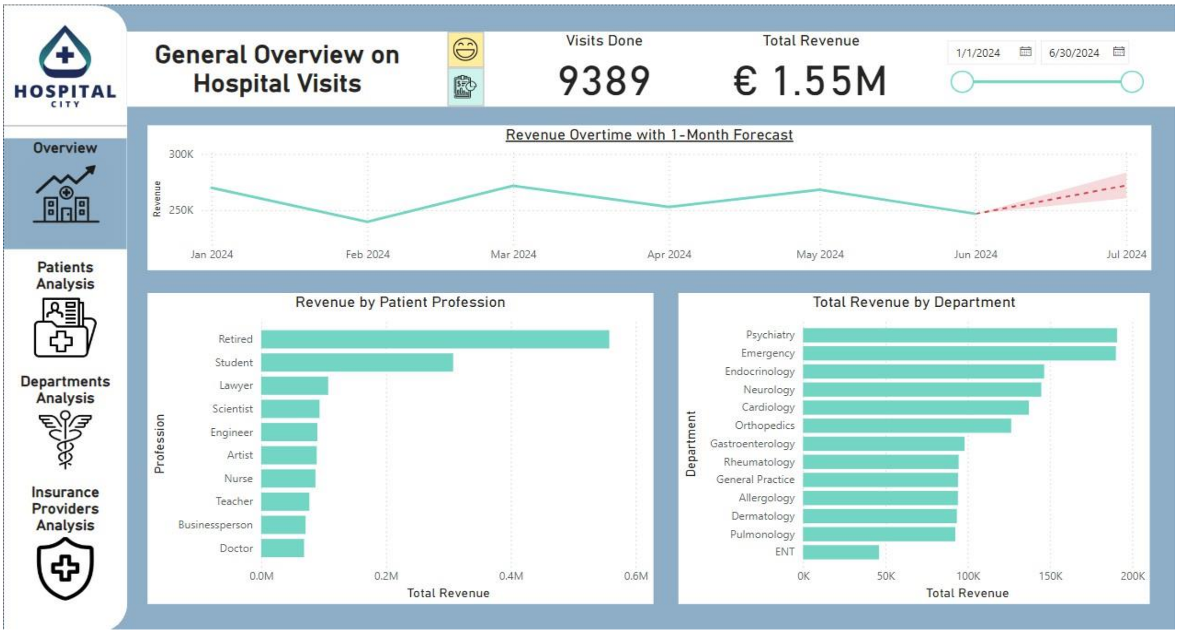


(*figure 5.1.*, data integration scheme on PowerBI)

PowerBI Dashboard Structure

We have selected the following themes for our visualizations.

General Overview: Firstly, we decided to give a general overview on the patient's visit. We split this section in two main parts; one part is focused on the financial situation of the hospital, the other on the satisfaction level. For each part we visualized the average overtime through a line chart and main distributions via histograms or pie charts. This section uses the transactional data only.



(*figure 5.2.*, general overview on financial situation)

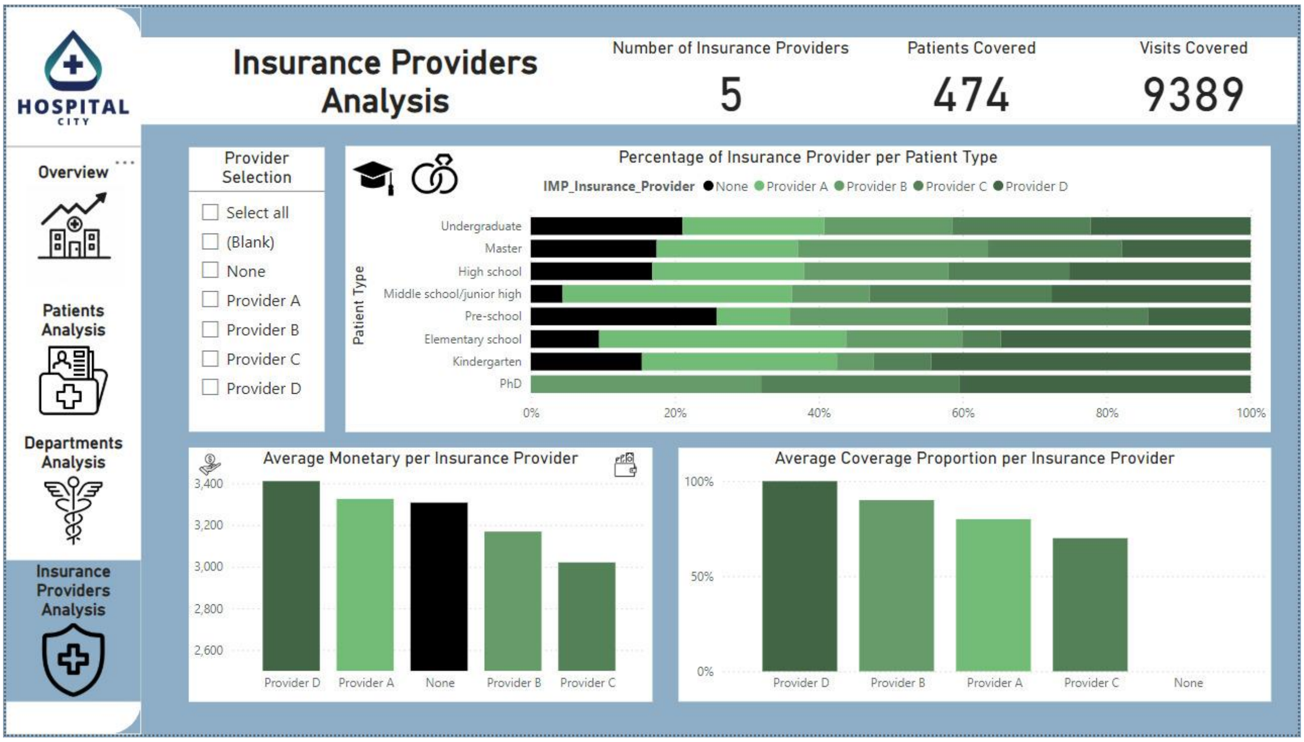
Patients Analysis: The best way to leverage the freshly constructed ABT table is to make a separate analysis for the patients, understanding its main patterns and distributions. We have decided to visualize the following: distribution of the patients by profession and gender, by their marital status (via 100%-histograms and donut charts), distribution of the frequency, membership and recency via line plots, and a scatter plot to visualize the correlation between age and income. Lastly, we also plotted a map of the patient's city of residences, to visualize their geographical zone.

(*figure 5.3.*, patients' analysis dashboard)

Departments Analysis: Another interesting theme to analyze is the hospital departments. In particular: we have plotted their frequencies over time via multiple line plots, average and variation of their consultation prices and the minimum-maximum price ranges of their consultation prices to understand the patterns in their prices. Lastly, we also plotted the average coverage proportion per department, to see which departments are covered the best.

(*figure 5.4.*, departments analysis dashboard)

Insurance Providers Analysis: Lastly, we also analyzed the insurance providers. We investigated the relationship between insurance providers and patients: to do this, we plotted various bar charts revealing the relationship between patient types and insurance provider types. Lastly, we plotted the average of insurance coverage proportion per insurance provider type, to understand the amount of price covered by each insurance provider. This view makes the best of both tables created during the preprocessing phase.



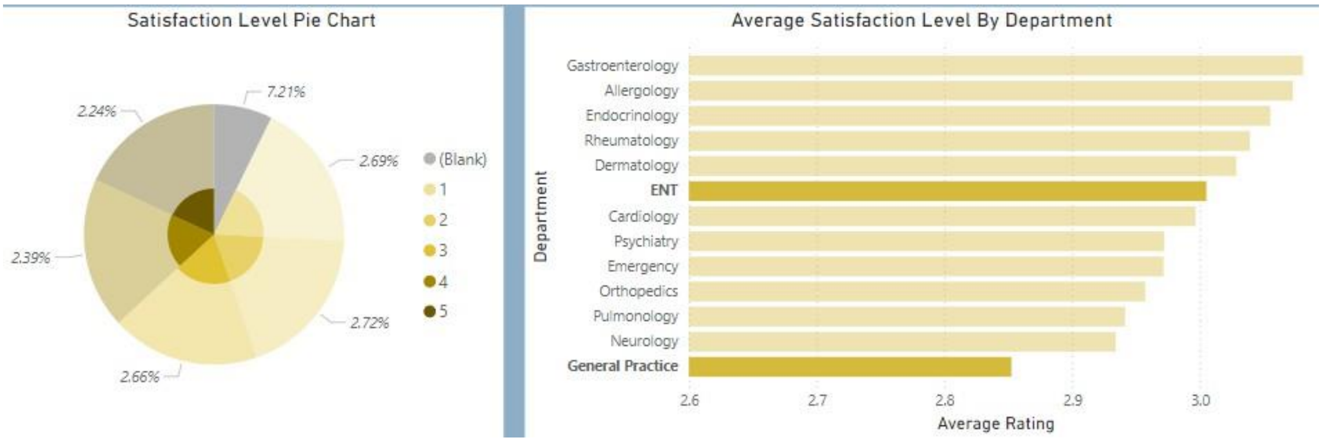
(*figure 5.5.*, insurance providers analysis dashboard)

Main Insights from the Visualizations

General Overview. From the general overview we gained some basic, but not insignificant, insights.

- Only General Practice and ENT received non-answers. This could be mainly since the patients feel that such types of visits were some sort of routine, therefore the patients felt that they were not compelled to give a rating; this might be especially the case for General Practice. In fact, other departments, which are more specific and critical, have ratings.
- General Practice seems to have the worst satisfaction level with an average of around 2.85; this could be due to the previous issue, that is a lack of ratings; therefore, this result could be not totally representative of the patients' satisfaction level. A recommendation would be to do targeted surveys towards patients of General Practice, to get deeper results.
- Psychiatry and Emergency are the best departments in terms of revenue, totaling to around 380K pounds, which is around the 25% of the total revenue.
- The financial situation is oscillating at around 64.5K pounds earned, by each month. The periodic nature could be due to the fact that visits are made in response to the season; for example, flu season could drastically increase the

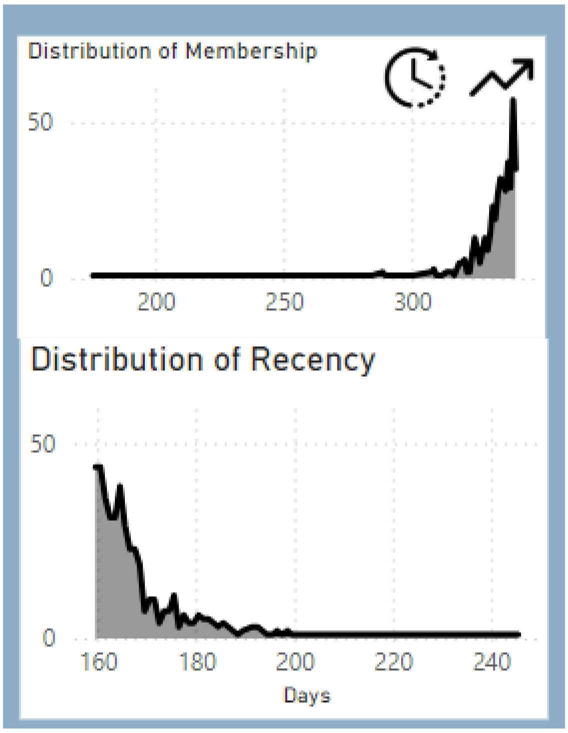
amount of visits, bringing more revenue.



(*figure 5.6.*)

Patients Analysis

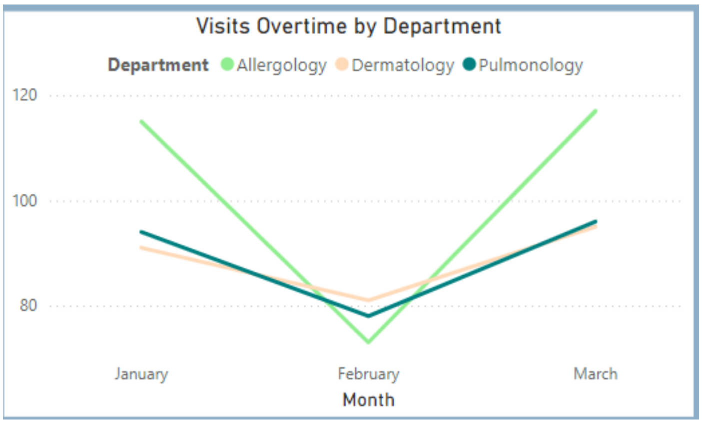
- All patients' city of residence seems to be from the British Isles, with a particular focus on the British patients as they make up about the 80.80% of the total patients.
- Some professions are dominated by a certain gender; Businesspersons and teachers are mostly female, whereas doctors are mostly male.
- There is a particular correlation between age and annual income; there is a cut-off at age 18, where underaged people have no income and all the others have a quasi-constant income. This is since every patient under 18 is still a student, thus has no income. Moreover, some middle-aged patients (age between 20-30) have a particularly high income from the rest of patients; they could be patients with professional positions.
- The frequency distribution resembles a normal distribution, with the mean frequency averaging to around 20. This means most of the patients had around 20 visits in the hospital. Moreover, the distribution of the patients' recency and membership seem to be closely correlated, as both are skewed distributions. From this we can see that most patients are "recent and loyal customers".



(*figure 5.7.*, distribution of Membership and Recency)

Departments Analysis

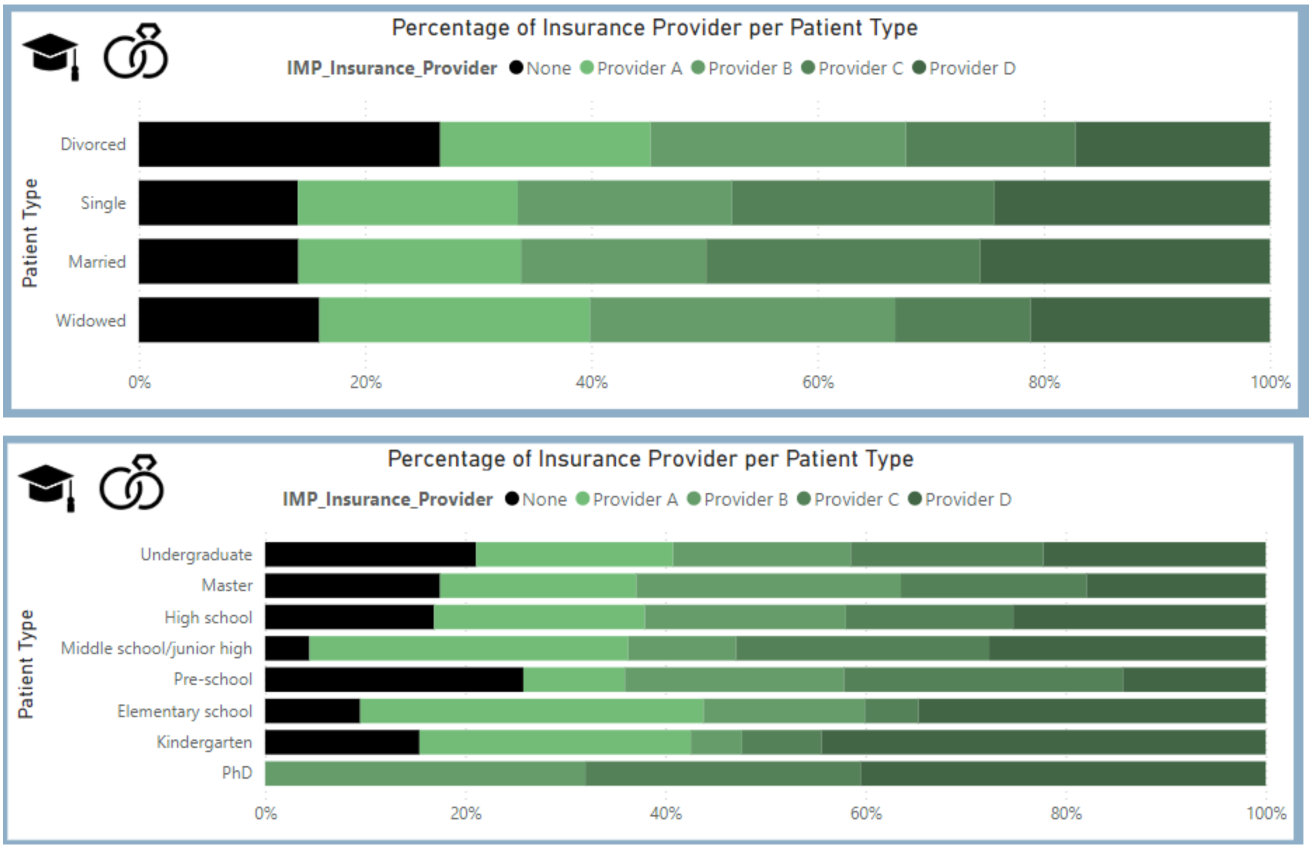
- Some departments experience a spike in visits between the months of February and March; they are Allergology, Dermatology and Pulmonology. This is mainly due to the *flu* and *allergy season*, as most of their symptoms require attention from the departments.
- Some departments have a generally higher number of visits than others, namely General Practice and Psychiatry. This is due to their "broadness" of the area, as they cover general problems, before referring them to specific departments.
- Each department's consultation cost is covered by from 67% to 72%. The lowest is Allergology with 67.35% coverage; this can be because their symptoms are not as "critical" as the others, so they tend to be covered the least by the insurance providers.
- Emergency has the highest statistical number in terms of average, variability (standard deviation) and minimum-maximum price range. This can be explained by the fact that emergencies are costly as they require critical resources and complex consultations and can cover a large range of critical problems.
- In the contrary, ENT and General practice has the lowest numbers in the previously mentioned terms. This is simply because they are usually simple consultations and require fewer resources.



(*figure 5.8.*, flu season spike)

Insurance Providers Analysis

- Most of patients having passed middle school/junior high school and having PhDs have insurance providers; regarding the PhDs, this could be because they have university-provided insurances, or they are required to have it as they work in critical environments (such as biological/chemical laboratories). Regarding the other education level, they could be required to have one as they are still underage.
- Divorced people have a higher rate of having no insurance providers in respect to others; we can explain it due to private health insurance dynamics, in the sense that insurance providers can be given as a spousal coverage. A divorce could cause a disruption in this and cause them to have no insurance in the meantime.
- Provider D seems to be the best insurance provider, as they cover - in average - the 99.75% of patients' consultation charges. Providers C, A, B seem to cover a decent number of charges, ranging from 70% to 90%. And of course, people with no insurance providers have no coverage at all.
- There might be a relationship between patients' choice of providers and their income or monetary: patients with insurance provider D tend to have a high average monetary value. This can be explained to be patients' strategic decision to choose a highly paying insurance provider as they are aware of their spendings on the hospital.


Percentage of Insurance Provider per Patient Type
IMP_Insurance_Provider ●None ●Provider A ●Provider B ●Provider C ●Provider D
Patient Type
Divorced
Single
Married
Widowed
0%
20%
40%
60%
80%
100%
Percentage of Insurance Provider per Patient Type
IMP_Insurance_Provider ●None ●Provider A ●Provider B ●Provider C ●Provider D
Patient Type
Undergraduate
Master
High school
Middle school/junior high
Pre-school
Elementary school
Kindergarten
PhD
0%
20%
40%
60%
80%
100%

(figure 5.9.)

Conclusion

The City Hospital project encompasses data preprocessing, Analytic Base Table (ABT) construction using SAS Studio, and data visualization with Power BI. This means the project represents a significant advancement in hospital consultations' analytics, as it lays the foundation for advanced analytics.

The team's project pipeline began with an exploratory data analysis (EDA) to ensure data integrity and address key challenges - such as outliers and missing values. This ensures that the given transactional data can be used for data analytics methods.

Subsequently, the preprocessing phase was followed by an enriched Analytic Base Table (ABT) focused on customer metrics, such as frequency, recency, and monetary value.

The cleaned transactional data and ABT laid the foundation for creating basic visualizations. In fact, the data preprocessing team also made interactive visualization dashboards, providing both an overview of transactions with basic insights into overall transaction patterns and focused analyses detailing patients, departmental operations, and insurance providers' data. This enables the hospital company to make data-driven strategic decisions.

In conclusion, the cleaned transactional table and ABT developed during this project hold a robust foundation for data mining operations. They include clustering (descriptive methods) for gaining insights on transactions or patients' patterns and predictive methods to predict important metrics such as revenue or rating of a customer.

Appendix

In the appendix we will report the SAS code used to perform data modification, ranging from checking for data inconsistencies to constructing the ABT.

(*snippet A.1.*, code for checking data consistency)

```
/* Program to check for basic consistency in the transactional table,
inconsistencies end up in deletion */

DATA CONSISTENT_TRANTABLE;
SET WORK.PREABT; /* File import */

/* Age has to be >0 */
IF (IMP_Age<0 OR IMP_Age=0) THEN DO;
    DELETE;
END;
/* 9867 -> 9724 */

/* Legal age for marriage in UK is 18, so any rows not respecting this is
considered as an inncosistency */
```

```
IF (IMP_Age<18 AND NOT(Marital_Status='Single')) THEN DO;
    DELETE;
END;
/* 9724 -> 9719 */

/* School leaving age is legally defined to be 16, therefore anyone with age
<=16 must be a student */
IF (IMP_Age<17 AND NOT(Profession='Student')) THEN DO;
    DELETE;
END;
/* 9719 -> 9610 */

/* Insurance coverage should be always smaller than consultation cost */
IF (Consultation_Price < IMP_Insurance_Coverage) THEN DO;
    DELETE;
END;
/* 9610 -> 9599*/

/* People without insurance should not have insurance coverage */
IF (IMP_INSURANCE_COVERAGE > 0 AND IMP_Insurance_Provider='None') THEN DO;
    DELETE;
END;
/* 9599 -> 9599 */

/* Check professions according to their degree required
    Lawyer, Engineer, Scientist -> At least high school
    Others won't be checked as some of them might have more specific
requirements
*/
IF (
    (PROFESSION='ENGINEER' OR PROFESSION='Lawyer' or PROFESSION='Scientist')
AND
    NOT(IMP_Education_Level='PhD' or IMP_Education_Level='Master' or
    IMP_Education_Level='Undergraduate' or IMP_Education_Level='High school')
) THEN DO;
    DELETE;
END;
/* 9599 -> 9599 */

/* Students should not have an income (we will not count cases of part-time
jobs or irregular work) */
IF (PROFESSION='Student' AND IMP_Approximate_Annual_Income > 0) THEN DO;
    DELETE;
END;
/* 9599 -> 9599 */

/* Compare age with education level
    High School: must be at least 16, compulsory education ends at that age
    Undergraduate: must be at least 21 (three years to complete a BsC degree)
```

```
    Master's: must be at least 22 (in UK master's last one year)
    PhD: 25 (3 years)
The rest won't be checked as the cases can vary. A discrepancy tolerance is
implemented.
*/
IF ( (IMP_EDUCATION_LEVEL='High school' AND IMP_AGE < 15 ) OR
     (IMP_EDUCATION_LEVEL='Undergraduate' AND IMP_AGE < 20) OR
     (IMP_EDUCATION_LEVEL='Master' AND IMP_AGE < 21) OR
     (IMP_EDUCATION_LEVEL='PhD' AND IMP_AGE < 24 )
) THEN DO;
    DELETE;
END;
/* 9599 -> 9389 */

/* People who have a paying job should have an income */
IF NOT(PROFESSION='Student' or PROFESSION='Retired') AND
IMP_APPROXIMATE_ANNUAL_INCOME=0 THEN DO;
    DELETE;
END;

/* 9389 -> 9389 */

/*
    RESULTS
    -------
    10 Queries
    9867 -> 9389 rows
    478 deleted rows
*/
```

(*snippet A.2.*, code for checking patients' data consistency)

```
/* SQL Queries to find Patient Inconsistencies */

PROC SQL;
    SELECT DISTINCT T1.PATIENT_ID, T1.PROFESSION, T2.PROFESSION
    FROM WORK.PREABT T1 LEFT JOIN WORK.PREABT T2
        ON T1.PATIENT_ID = T2.PATIENT_ID
    WHERE
        T1.PROFESSION <> T2.PROFESSION;
RUN;

PROC SQL;
    SELECT DISTINCT T1.PATIENT_ID, T1.IMP_AGE, T2.IMP_AGE
    FROM WORK.PREABT T1 LEFT JOIN WORK.PREABT T2
        ON T1.PATIENT_ID = T2.PATIENT_ID
    WHERE
        T1.IMP_AGE <> T2.IMP_AGE;
RUN;
```

```
PROC SQL;
    SELECT DISTINCT T1.PATIENT_ID, T1.GENDER, T2.GENDER
    FROM WORK.PREABT T1 LEFT JOIN WORK.PREABT T2
        ON T1.PATIENT_ID = T2.PATIENT_ID
    WHERE
        T1.GENDER <> T2.GENDER;
RUN;

PROC SQL;
    SELECT DISTINCT T1.PATIENT_ID
    FROM WORK.PREABT T1 LEFT JOIN WORK.PREABT T2
        ON T1.PATIENT_ID = T2.PATIENT_ID
    WHERE
        T1.IMP_INSURANCE_PROVIDER <> T2.IMP_INSURANCE_PROVIDER;
RUN;

PROC SQL;
    SELECT DISTINCT T1.PATIENT_ID, T1.FAMILY_HISTORY, T2.FAMILY_HISTORY
    FROM WORK.PREABT T1 LEFT JOIN WORK.PREABT T2
        ON T1.PATIENT_ID = T2.PATIENT_ID
    WHERE
        T1.FAMILY_HISTORY <> T2.FAMILY_HISTORY;
RUN;

PROC SQL;
    SELECT DISTINCT T1.PATIENT_ID, T1.MARITAL_STATUS, T2.MARITAL_STATUS
    FROM WORK.PREABT T1 LEFT JOIN WORK.PREABT T2
        ON T1.PATIENT_ID = T2.PATIENT_ID
    WHERE
        T1.MARITAL_STATUS <> T2.MARITAL_STATUS;
RUN;

PROC SQL;
    SELECT DISTINCT T1.PATIENT_ID, T1.CITY_OF_RESIDENCE, T2.CITY_OF_RESIDENCE
    FROM WORK.PREABT T1 LEFT JOIN WORK.PREABT T2
        ON T1.PATIENT_ID = T2.PATIENT_ID
    WHERE
        T1.CITY_OF_RESIDENCE <> T2.CITY_OF_RESIDENCE;
RUN;
```

(*snippet A.3.*, code for correcting patients' rows)

```
DATA CONSISTENT_TRANTABLE;
SET WORK.PREABT; /* File import */

IF ( PATIENT_ID=1488 AND NOT(PROFESSION='Student')) THEN DO;
        PROFESSION='Student';
END;
```

```
IF ( PATIENT_ID=1496 AND NOT(PROFESSION='Student')) THEN DO;
    PROFESSION='Student';
END;

IF ( PATIENT_ID=1050 AND NOT(GENDER='Female')) THEN DO;
    GENDER='Female';
END;

IF ( PATIENT_ID=1307 AND NOT(GENDER='Male')) THEN DO;
    GENDER='Male';
END;

IF ( PATIENT_ID=1349 AND NOT(GENDER='Male')) THEN DO;
    GENDER='Male';
END;

IF ( PATIENT_ID=1447 AND NOT(GENDER='Female')) THEN DO;
    GENDER='Female';
END;

IF ( PATIENT_ID=1490 AND NOT(GENDER='Female')) THEN DO;
    GENDER='Female';
END;

IF ( PATIENT_ID=1140 AND NOT(MARITAL_STATUS='Single')) THEN DO;
    MARITAL_STATUS='Single';
END;

IF ( PATIENT_ID=1322 AND NOT(MARITAL_STATUS='Single')) THEN DO;
    MARITAL_STATUS='Single';
END;

IF ( PATIENT_ID=1332 AND NOT(MARITAL_STATUS='Single')) THEN DO;
    MARITAL_STATUS='Single';
END;

IF ( PATIENT_ID=1382 AND NOT(MARITAL_STATUS='Single')) THEN DO;
    MARITAL_STATUS='Single';
END;
```

(*snippet A.4.*, ABT creation code)

```
PROC SQL;
CREATE TABLE BIO_INFO AS
    SELECT DISTINCT PATIENT_ID, GENDER, PROFESSION, FAMILY_HISTORY,
CITY_OF_RESIDENCE, MARITAL_STATUS, IMP_INSURANCE_PROVIDER
    FROM WORK.PREABTCONSISTENT /* IMPORTANT !!! */
    GROUP BY PATIENT_ID;
RUN;
/* ^^ Directly transposes some biographical/anagraphical information ^^ */
```

```
    /* such as gender, profession, family history, which are supposed to be
unique. */

/* ======================================================== */
CREATE TABLE AGE AS
    SELECT DISTINCT PATIENT_ID, avg(IMP_Age) as Age
    FROM WORK.PREABTCONSISTENT
    GROUP BY PATIENT_ID;
RUN;
/* As there are inconsitencies in the imputed ages, we will simply take their
average */
/* ======================================================== */

PROC SQL;
CREATE TABLE STEP1 AS
    SELECT X.PATIENT_ID, X.DEPARTMENT, (sum(X.Consultation_Price)/T.MON) as
TotAmt
    FROM WORK.PREABTCONSISTENT as X, (
        SELECT PATIENT_ID, sum(AUX.CONSULTATION_PRICE) as MON
        FROM WORK.PREABTCONSISTENT AS AUX
        GROUP BY AUX.PATIENT_ID) as T
    WHERE T.PATIENT_ID = X.PATIENT_ID
    GROUP BY X.PATIENT_ID, X.DEPARTMENT;
RUN;

PROC SORT DATA=STEP1 OUT=STEP2;
    BY PATIENT_ID;
RUN;

PROC TRANSPOSE DATA=STEP2 OUT=SEGMENTED_PRICE
    PREFIX=proportion_price_;
    ID DEPARTMENT;
    BY PATIENT_ID;
RUN;

/* ^^ Segments total consultation price by department in form of proportion ^^
*/

/* ======================================================== */
PROC SQL;
CREATE TABLE STEP1 AS
    SELECT PATIENT_ID, DEPARTMENT, count(*) as Freq
    FROM WORK.PREABTCONSISTENT
    GROUP BY PATIENT_ID, DEPARTMENT;
RUN;

PROC SORT DATA=STEP1 OUT=STEP2;
    BY PATIENT_ID;
RUN;
```

```
PROC TRANSPOSE DATA=STEP2 OUT=SEGMENTED_FREQ
    PREFIX=freq_;
    ID DEPARTMENT;
    BY PATIENT_ID;
RUN;
/* ^^ same as above but with frequency */

/* ======================================================== */
proc sql;
CREATE TABLE MEMBERSHIP AS
    select distinct PATIENT_ID, (DATETIME()-min(Visit_Date))/86400 as
membership_days
    from WORK.PREABTCONSISTENT
    group by PATIENT_ID;
run;

/* ======================================================== */
proc sql;
CREATE TABLE RECENCY AS
    select distinct PATIENT_ID, (DATETIME()-max(Visit_Date))/86400 as
recency_days
    from WORK.PREABTCONSISTENT
    group by PATIENT_ID;
run;
/* Get recency */

/* ======================================================== */
PROC SQL;
CREATE TABLE AGGREGATED_INFO AS
    SELECT PATIENT_ID,
        avg(Consultation_Duration) as avg_duration,
        avg(Satisfaction_Level) as avg_Satisfaction_Level,
        sum(Consultation_Price) as monetary,
        avg(IMP_Approximate_Annual_Income) as avg_recorded_income,
        count(*) as total_frequency
    FROM WORK.PREABTCONSISTENT
    GROUP BY PATIENT_ID;
RUN;

PROC SQL;
CREATE TABLE SAT_LEV AS
    SELECT PATIENT_ID,
        avg(Satisfaction_Level) as avg_Satisfaction_Level
    FROM WORK.PREABTCONSISTENT
    WHERE Satisfaction_Level < 6
    GROUP BY PATIENT_ID;
RUN;
```

```
/* Get important aggregated variables*/
    /* Namely: -total amount of money spent; -mode of department; -
satisfaction, duration, ANI avg. */

/* ======================================================== */
PROC SQL;
CREATE TABLE PROPORTION_COVERAGE AS
    SELECT DISTINCT X.PATIENT_ID, sum(X.IMP_Insurance_Coverage)/T.MON as
CoverageProportion
    FROM WORK.PREABTCONSISTENT as X, (
        SELECT PATIENT_ID, sum(AUX.CONSULTATION_PRICE) as MON
        FROM WORK.PREABTCONSISTENT AS AUX
        GROUP BY AUX.PATIENT_ID) as T
    WHERE T.PATIENT_ID = X.PATIENT_ID
    GROUP BY X.PATIENT_ID
;
RUN;

/* ======================================================== */
DATA PRE_FINAL;
    MERGE BIO_INFO AGE PROPORTION_COVERAGE SEGMENTED_PRICE SEGMENTED_FREQ
RECENCY MEMBERSHIP SAT_LEV AGGREGATED_INFO;
    BY PATIENT_ID;
RUN;

DATA PRE_PRE_FINAL;
    SET PRE_FINAL;
    DROP _NAME_;
RUN;

DATA FINAL_ABT;
    SET PRE_PRE_FINAL;
    ARRAY change _numeric_;
        DO OVER change;
        IF change=. THEN change=0;
    END;
RUN;

/* Exclude patients with different insurance providers from ABT */
DATA FINAL_ABT;
    SET FINAL_ABT;
IF (
    PATIENT_ID in(
1013,
1014,
1015,
1028,
1031,
1034,
```

```
1089,
1092,
1100,
1105,
1135,
1143,
1234,
1245,
1248,
1260,
1261,
1266,
1285,
1294,
1302,
1308,
1317,
1340,
1343,
1381,
1449,
1455,
1485,
1490,
1498
)
)
THEN DO;
    DELETE;
END;
```