HEALTHCARE DATASET

---------------------------------------------------------------------------------------------------

# Business Problem:

# The healthcare sector is an indispensable market, intricately intertwined with individuals' well-being. It is imperative for businesses to proactively address the challenges within this domain. Financial dynamics play a pivotal role in the equation as medical treatments can often impose substantial financial burdens. Individuals who lack adequate insurance coverage may find themselves in precarious financial situations.

# Insurance companies operating in the healthcare space are faced with the pressing need to optimize their insurance offerings. This optimization is not solely for profitability but is also driven by the desire to mitigate risk. A healthy lifestyle and responsible health choices, such as proper nutrition and exercise, significantly reduce the likelihood of illness. Thus, insurance providers aim to promote and incentivize these behaviors among policyholders to curtail overall healthcare costs and enhance the long-term sustainability of their offerings.

# In this report, we delve into the intricate relationship between healthcare, insurance, and individual well-being. We explore how the convergence of these factors presents both challenges and opportunities in the healthcare industry.

# Importing the Dataset.

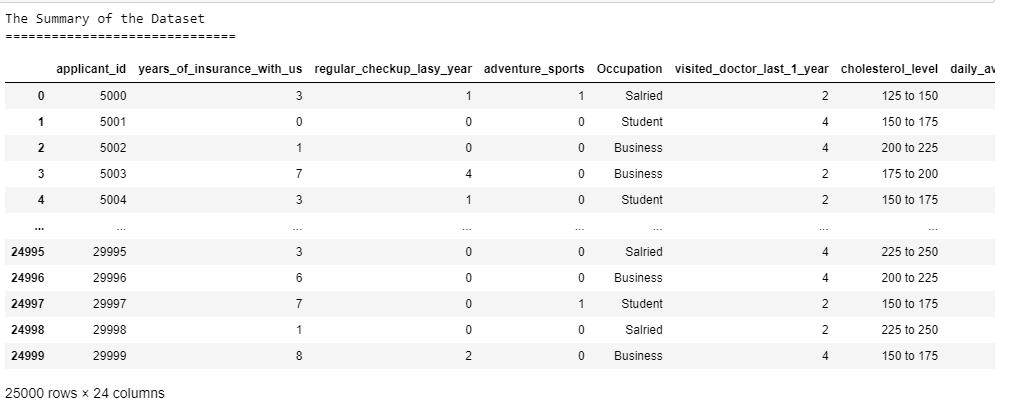


Table no 1: Summary of the dataset

# Head and Tail of the dataset:

### 

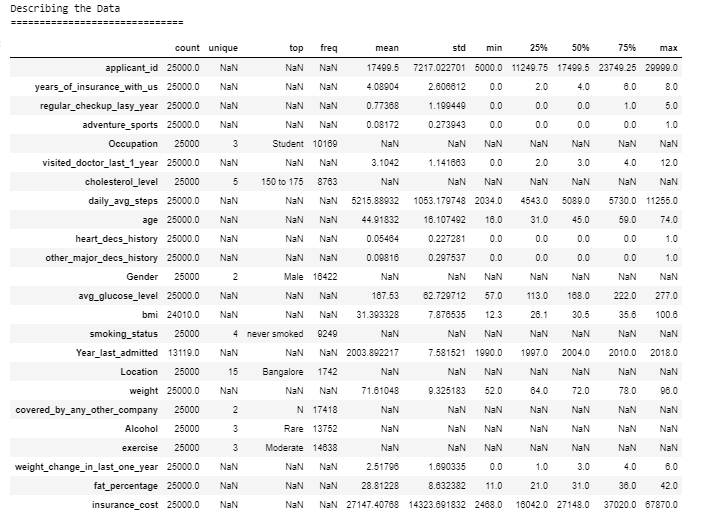
### 

### Table no 2: Head and Tail of the dataset

### Inferences:

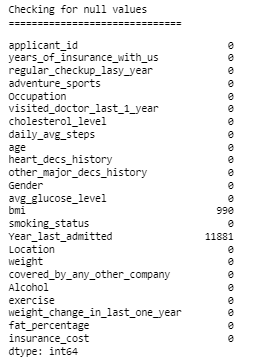
* The first ten rows of the dataset which is the head of the dataset and the last ten rows of the dataset which is the tail of the dataset.
* There are a total of 25000 entries in the dataset.
* The dataset has 25000 rows and 24 columns.

# Describing the dataset:



### Table no 3: Description of the dataset

# Checking for Null Values



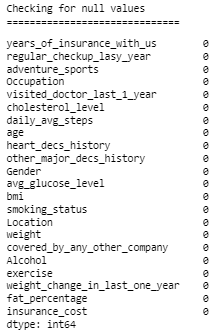


Table no 4 : Sum of Null Values

# Checking for Duplicate Values



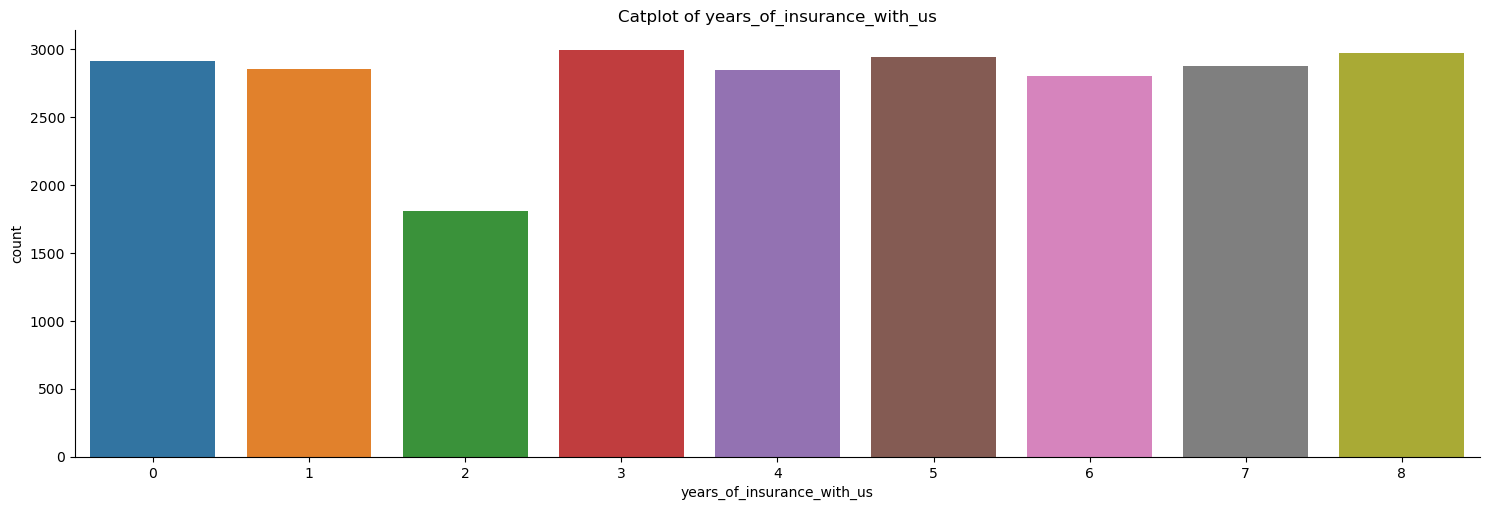
Table no 5: Sum of duplicate values

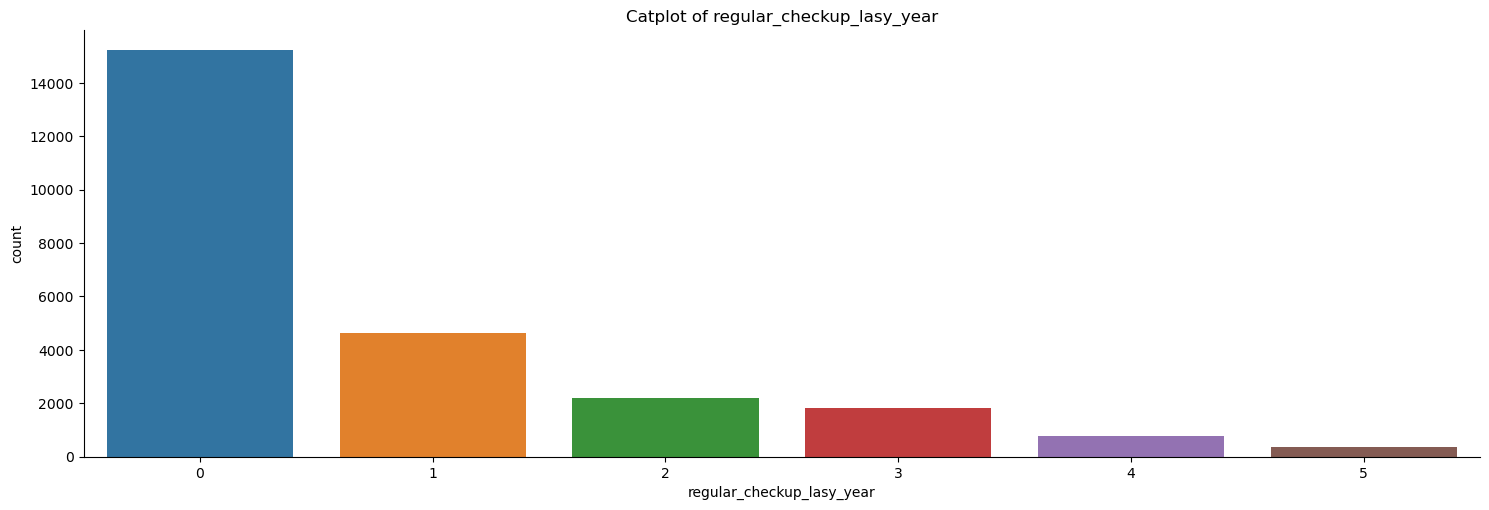
### Inferences:

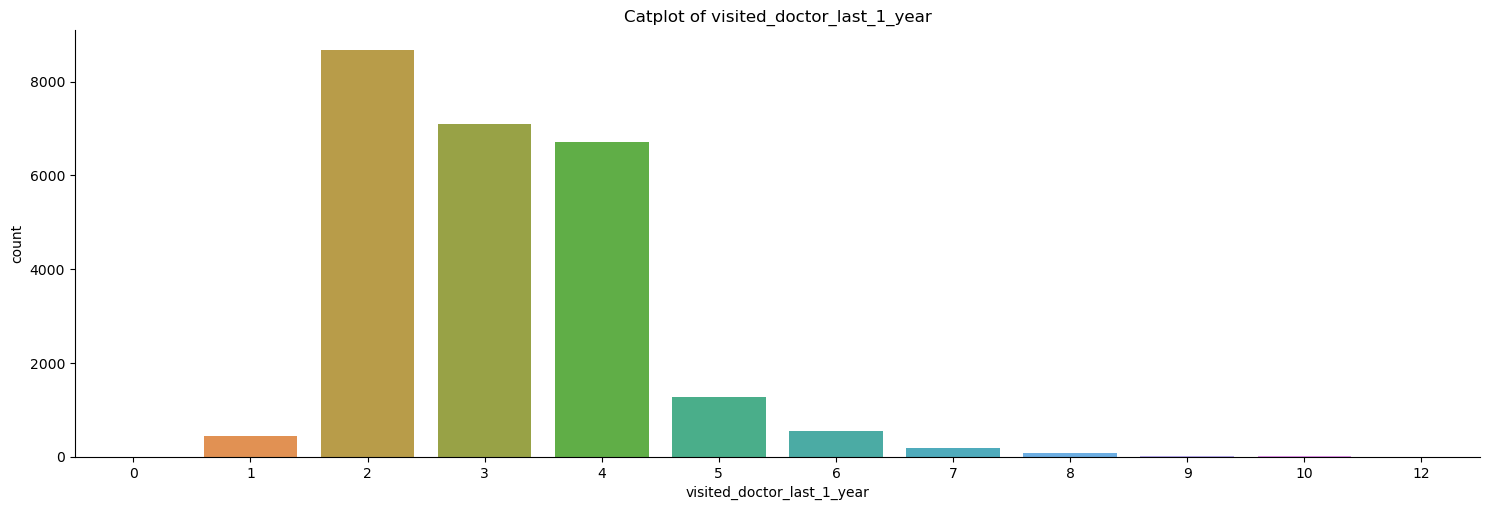
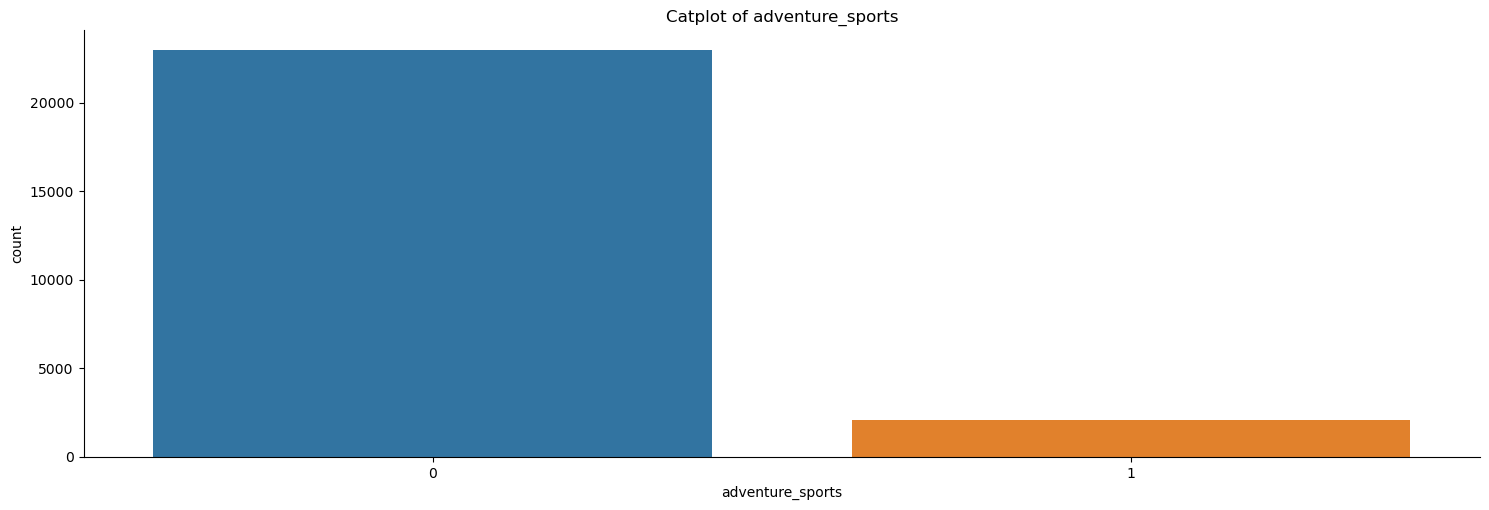
* There are no duplicate values in the null values in the dataset, namely 990 in the column ‘bmi’ and 11001 in the column ‘Year\_last admitted’.

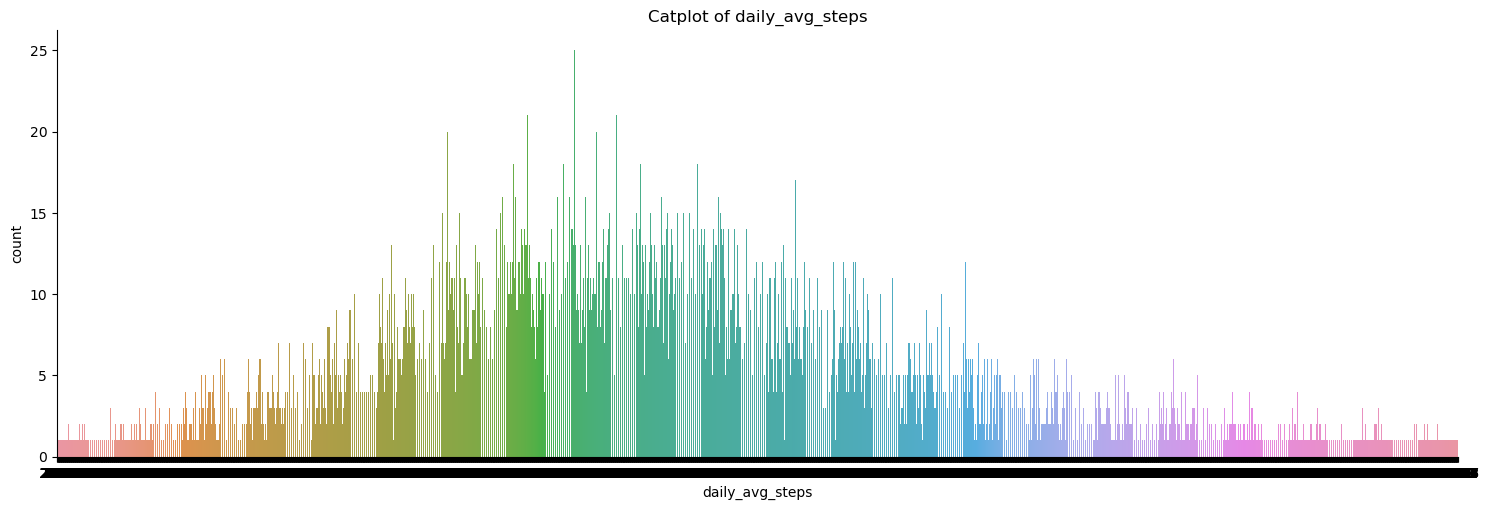
# Univariate Analysis:

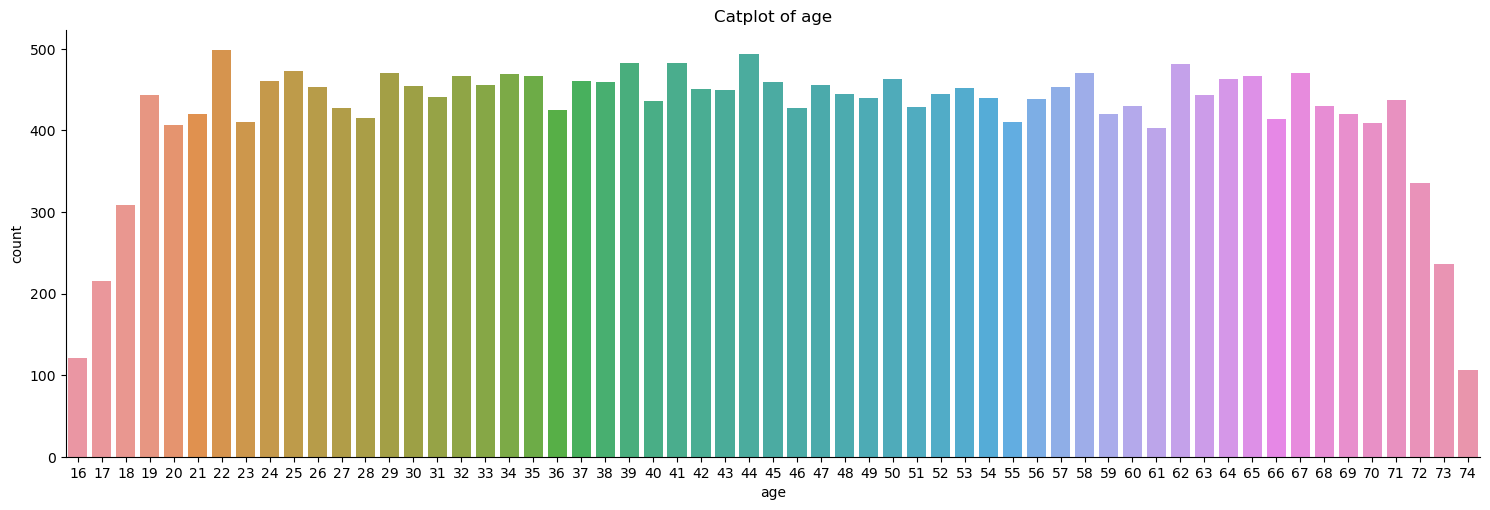
## CATPLOT

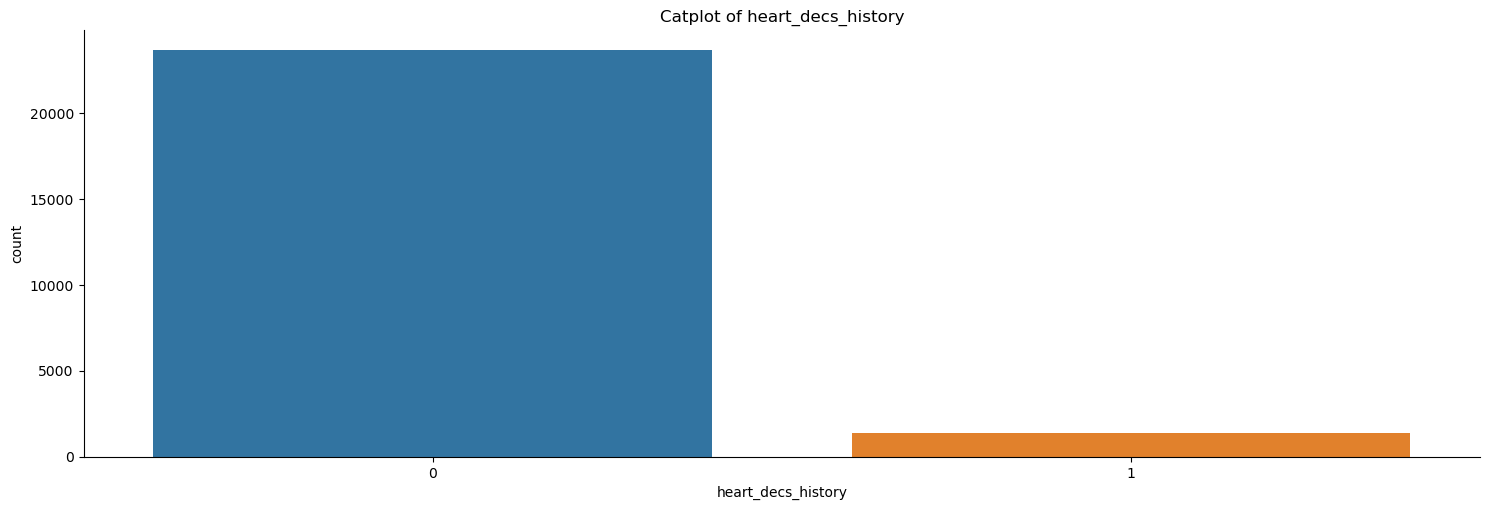


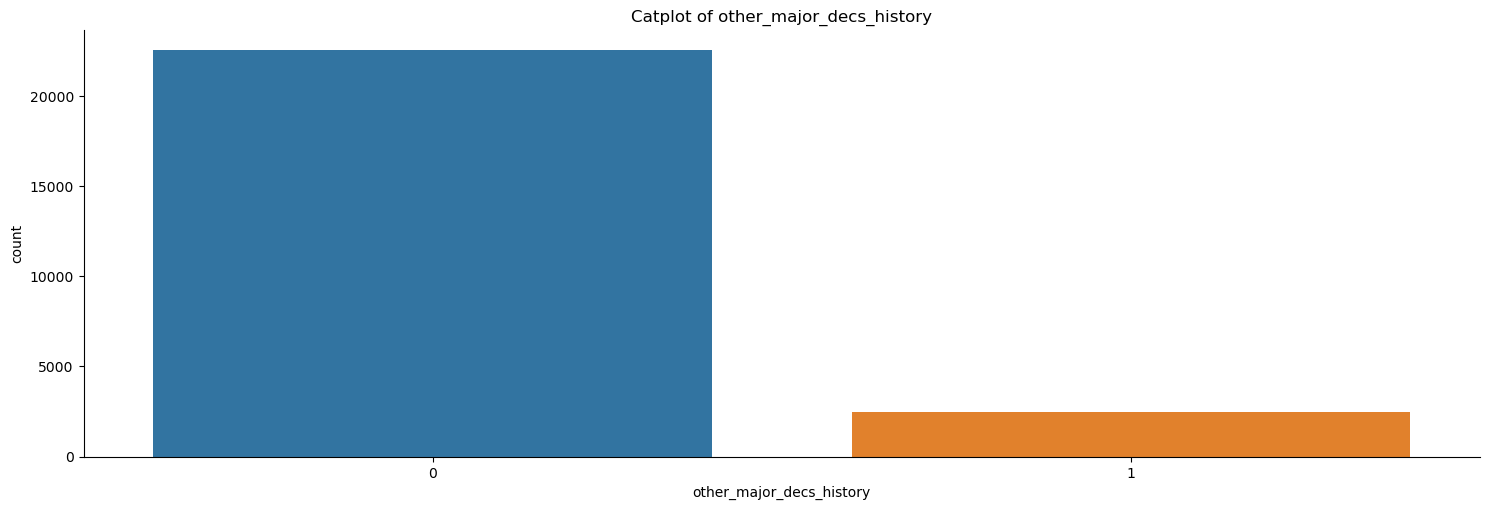


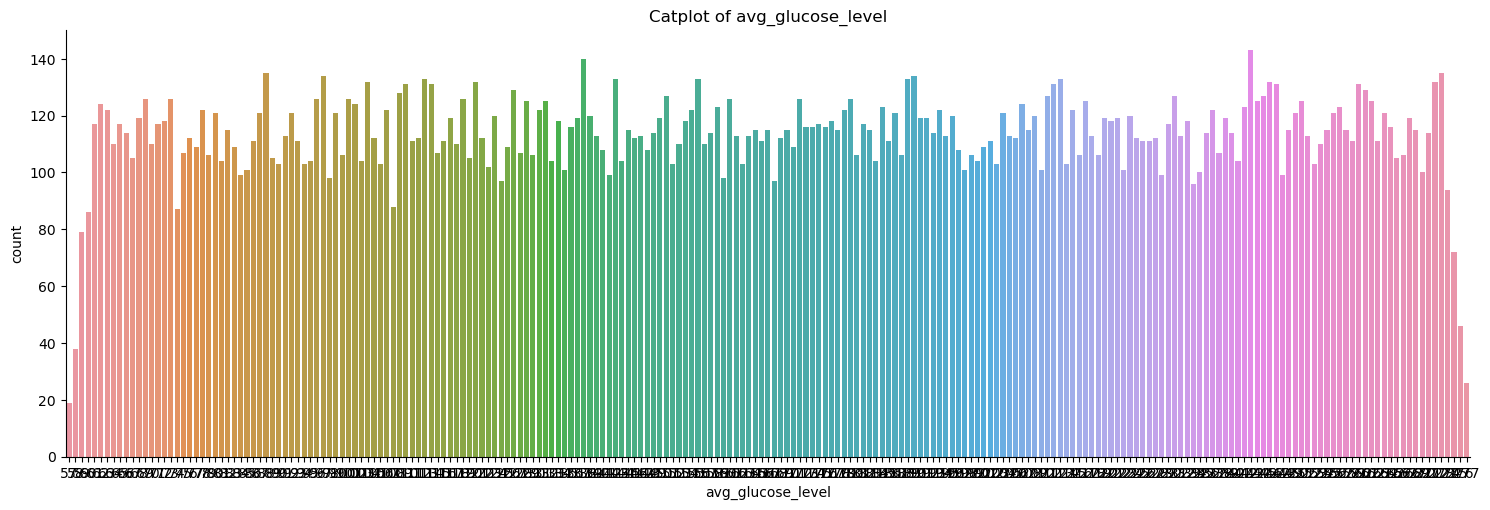


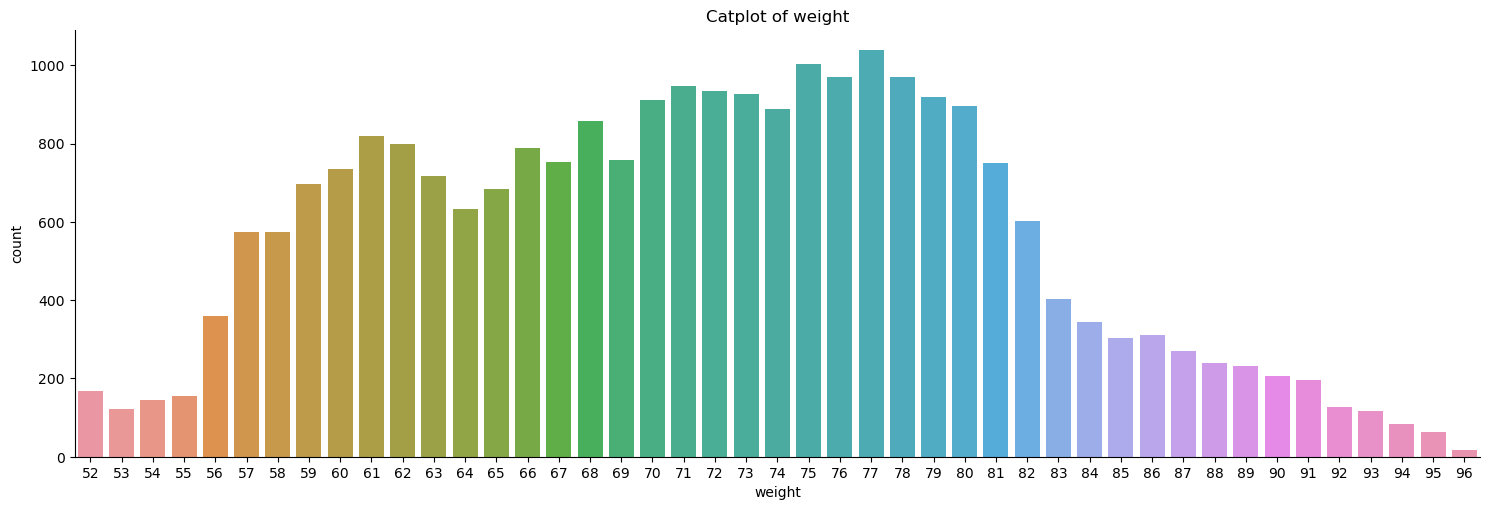


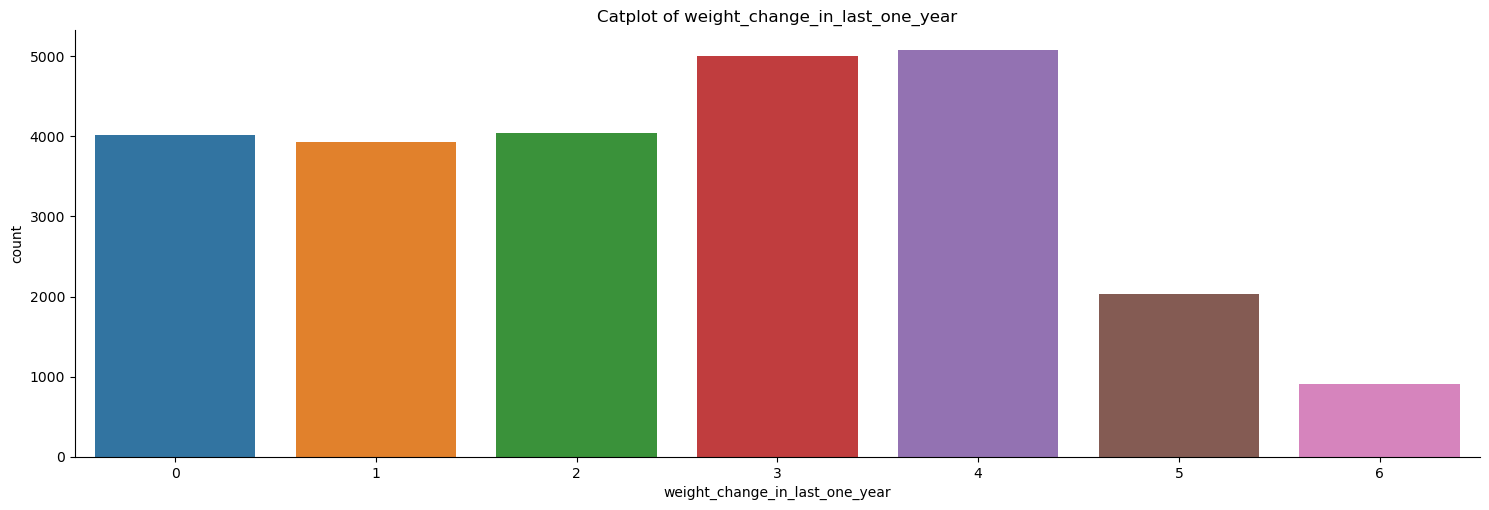


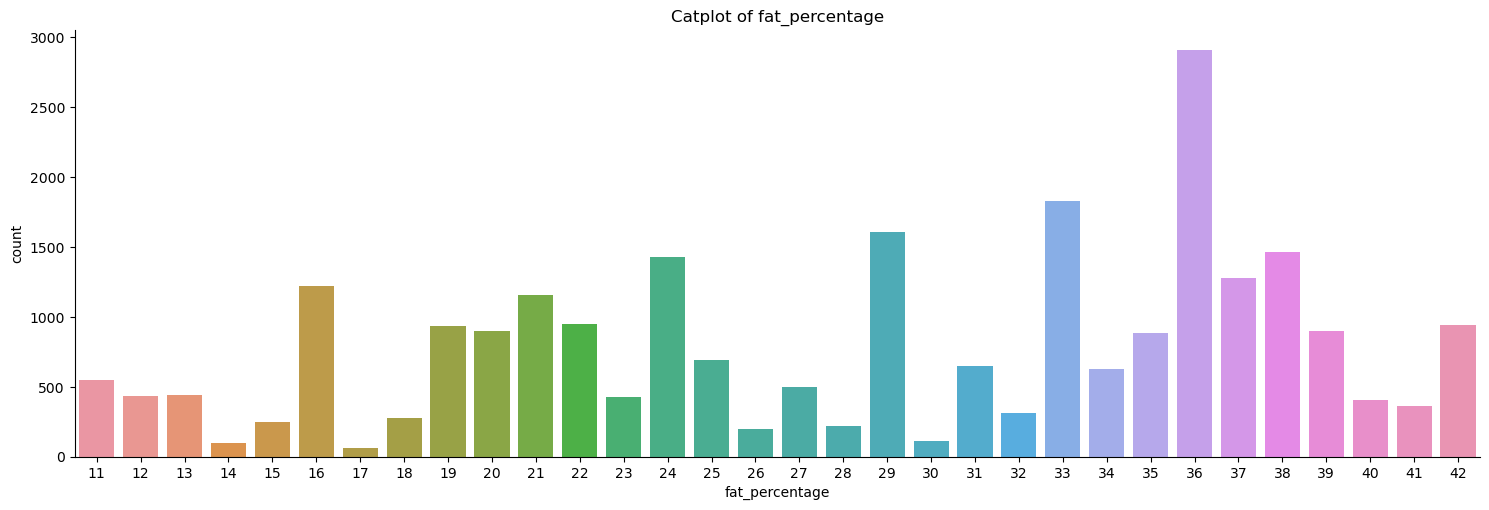


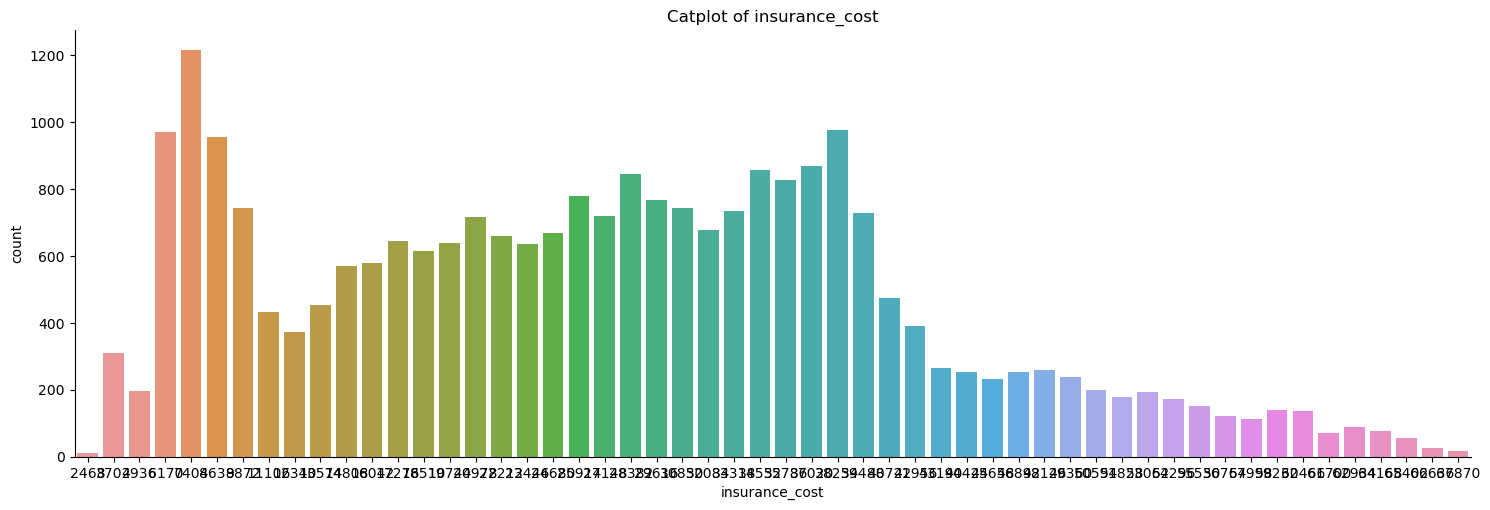


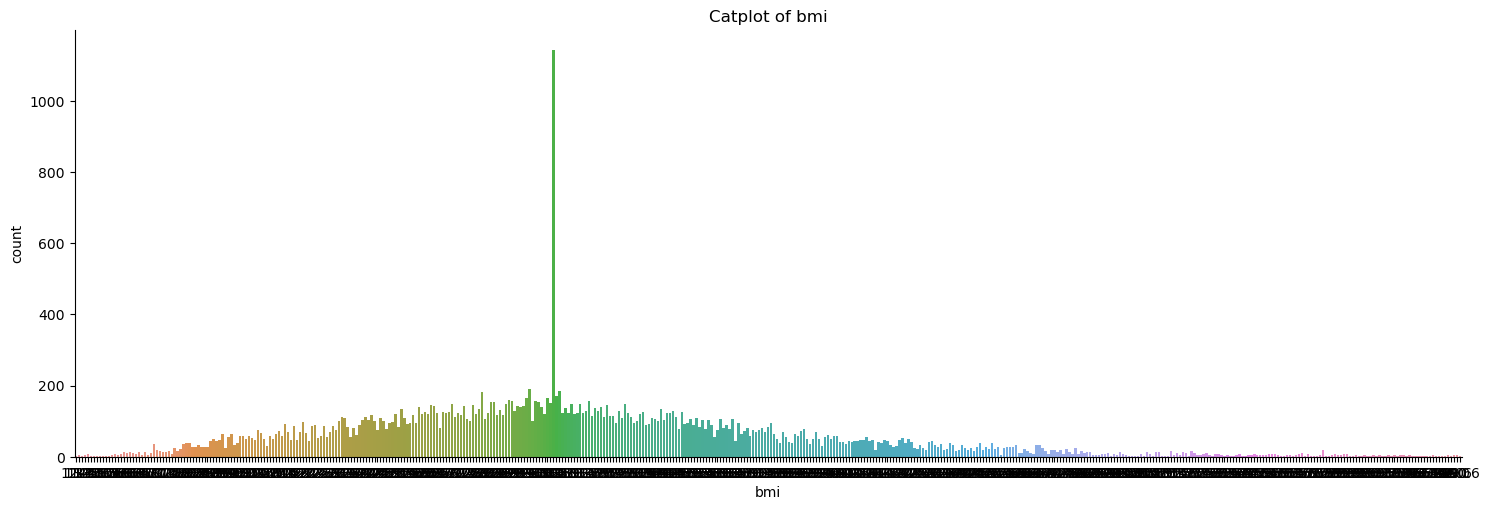




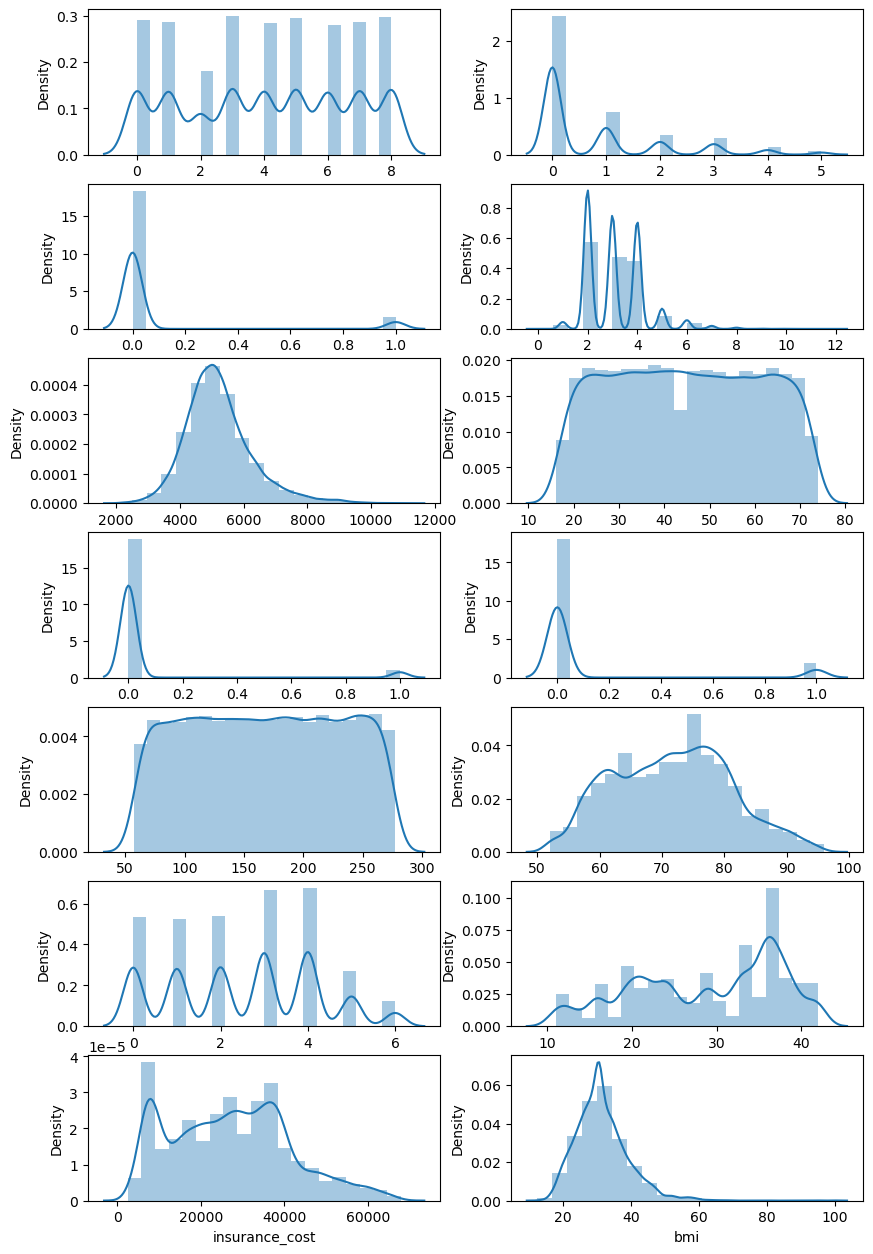




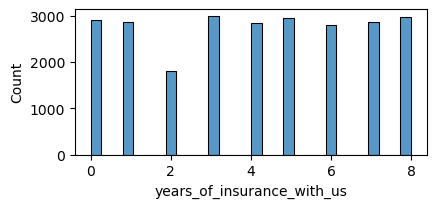


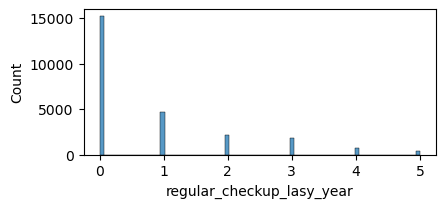


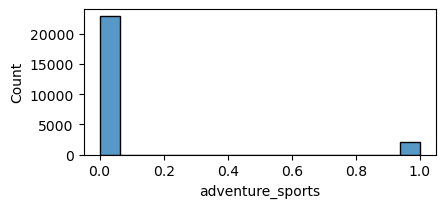
## Distplot

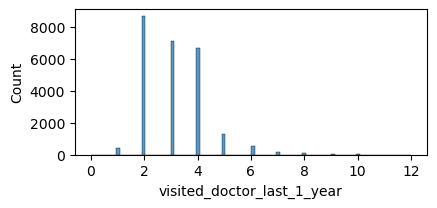


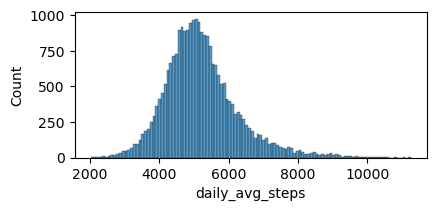
## HISTPLOT

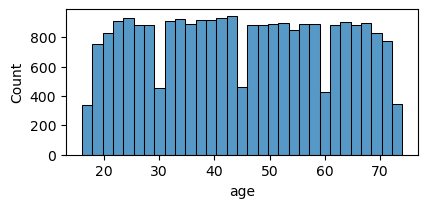




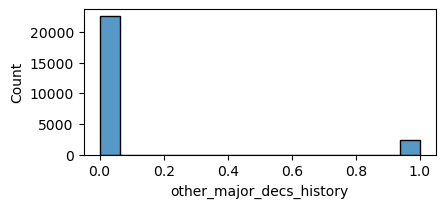


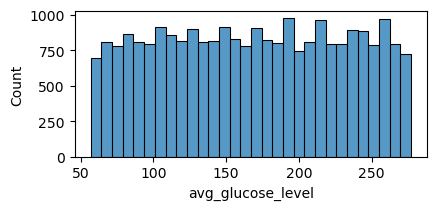


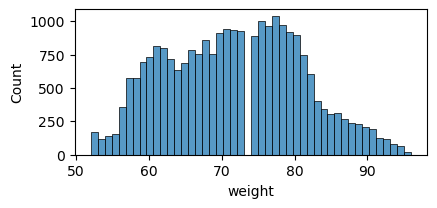


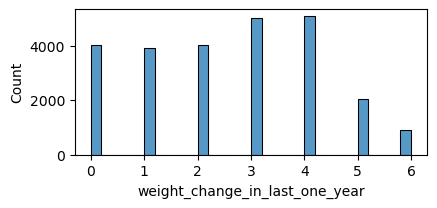


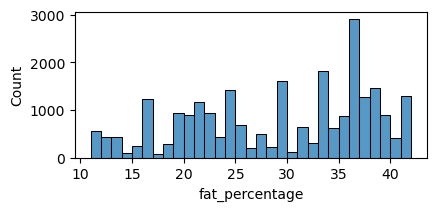


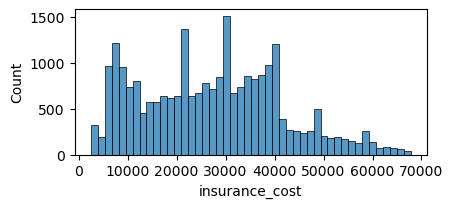


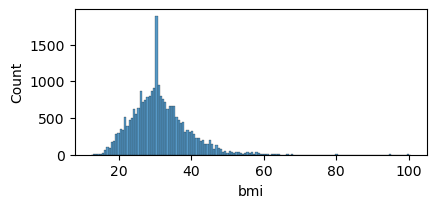




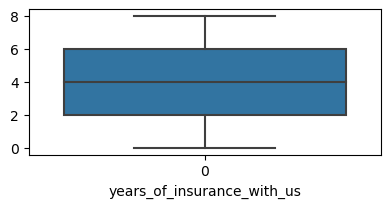


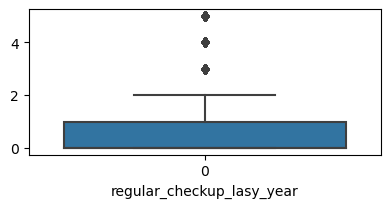


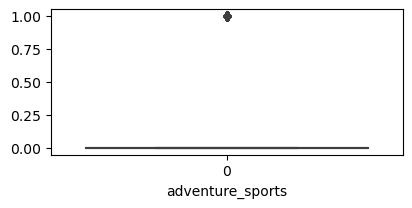


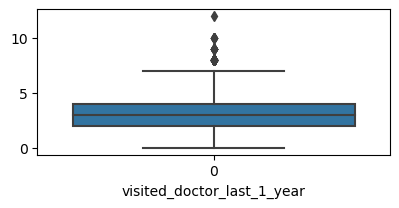


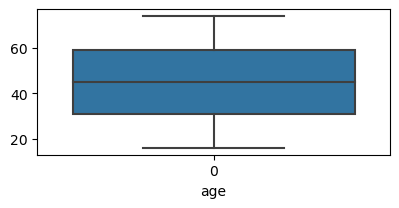
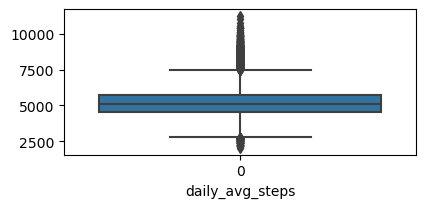
## BOXPLOT

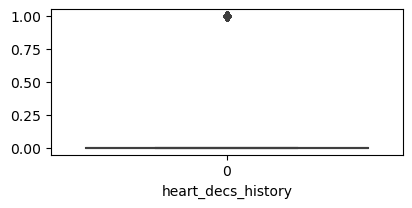


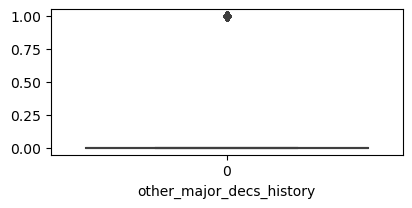


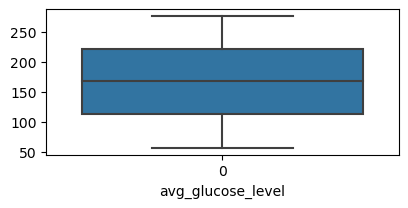


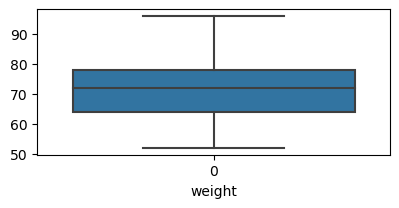


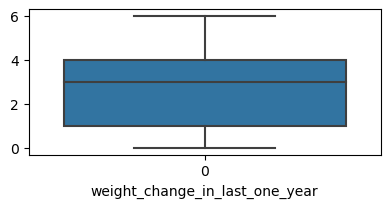


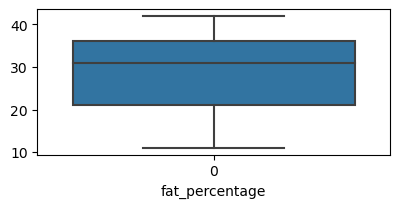


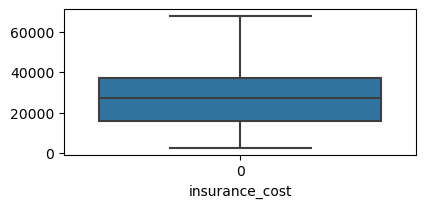


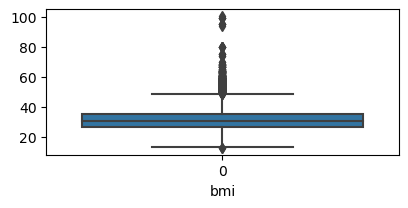












### Inferences:

* The dataset contains outliers.

# Skew

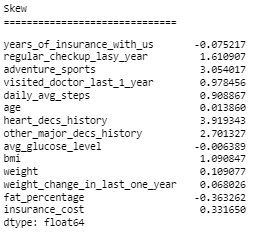


Table no 11: Skew

# Multivariate Analysis:

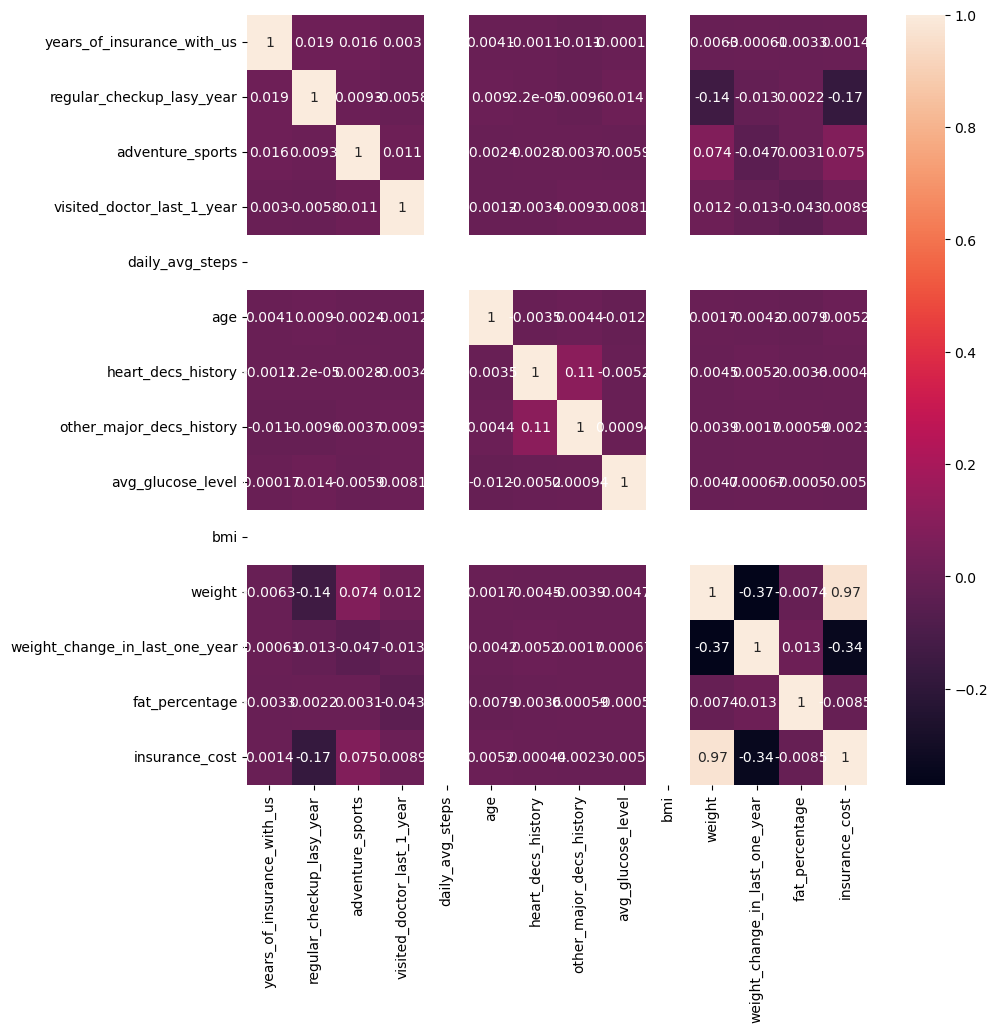


Table no 13: Heatmap

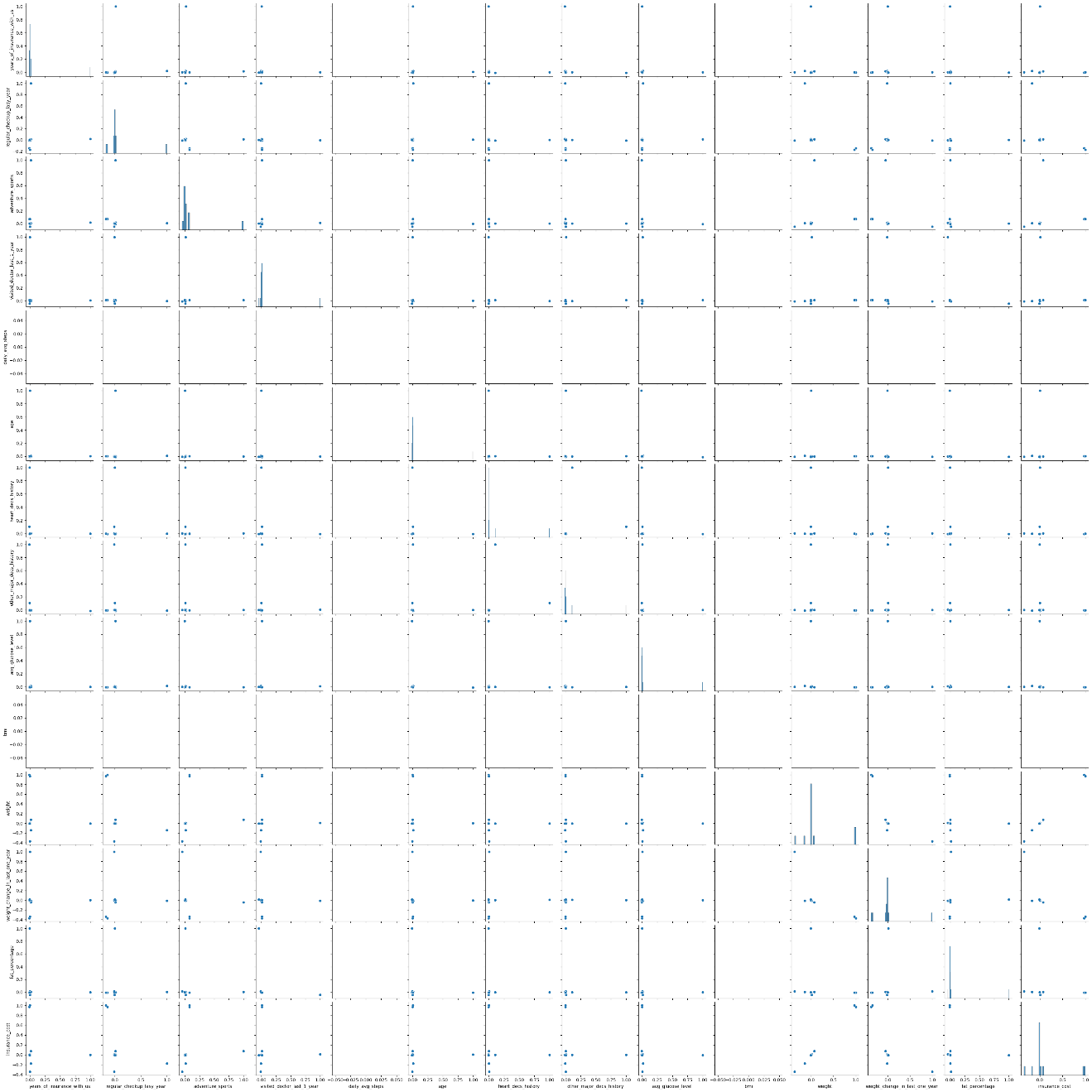


Table no 14: Pairplot

# MODEL BUILDING

## decision tree

### Checking model score





* The model is overfitted.

### Pruning the model

Best params:



### Best model score





## RANDOM FOREST

### Checking model score

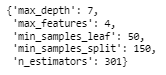




* The model is overfitted.

### Pruning the model

Best params:



### Best model score





## LINIEAR REGRESSION

### Checking model score



* The model is overfitted.

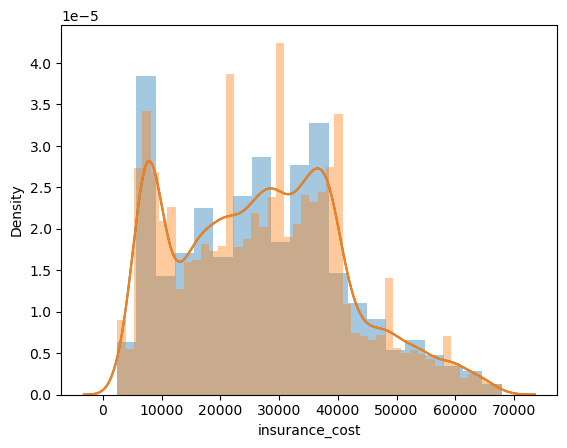
### Pruning the model

* The variables Alcohol, Avg glucose level, Smoking status, Gender Male, Exercise, Location, Other major decs history, and Occupation have been removed.

### Best model score



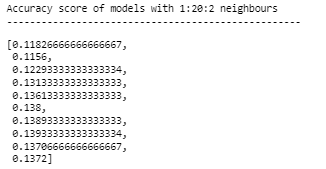
Distplot of prediction :

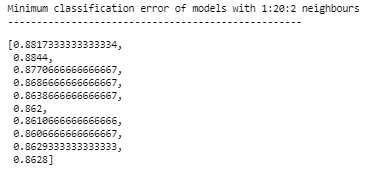


## KNN

### Checking model score

### Best model score





* The model with 15 neighbors is the best model.

# MODEL BUILDING WITH THE SCALED DATASET

## decision tree

### Checking model score

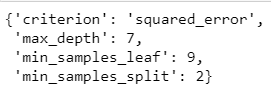




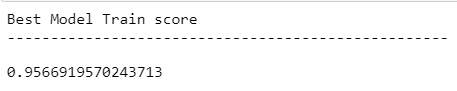
* The model is overfitted.

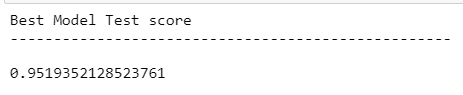
### Pruning the model using GridSearchCV

Best params:



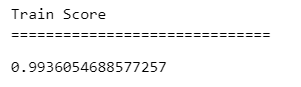
### Best model score

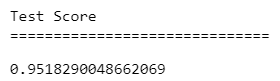




## RANDOM FOREST

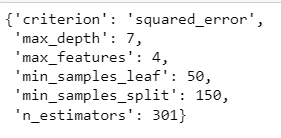
### Checking model score



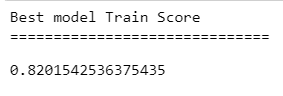


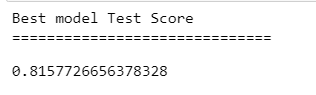
### Pruning the model using GridSearchCV

Best params:



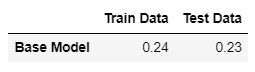
### Best model score





## LINIEAR REGRESSION

### Checking model score

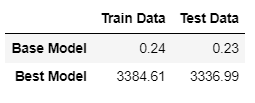


* The model is overfitted.

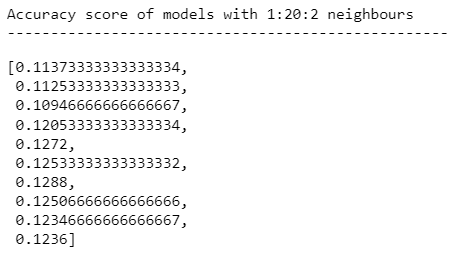
### Pruning the model

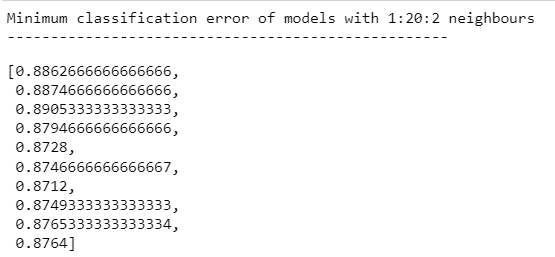
* The variables Alcohol, Avg glucose level, Smoking status, Gender Male, Exercise, BMI, fat percentage, avg steps, cholesterol level, years of insurance with us, Other major decs history, and Occupation have been removed.

### Best model score



## KNN





* 13 neighbours is the best model.

