

MAHENDRAN N | EXCEL | November 30, 2024

EXCEL CAPSTONE PROJECT

EXCEL FEATURES

**Data Cleaning:**

1. Check for the number of missing values marked with '?' in each column of the “Medical Examinations” Table and "Hospitalization Details" Table.

We can find the missing value marked “?” by using COUNTIF function,

We have observed the below mentioned missing values in the given format.

### **Medical Examinations Table:**

* Customer ID: 0
* BMI: 0
* HBA1C: 0
* Heart Issues: 0
* Any Transplants: 0
* Cancer history: 0
* NumberOfMajorSurgeries: 0
* smoker: 2

### **Hospitalization Details Table:**

* Customer ID: 6
* year: 2
* month: 3
* date: 0
* children: 0
* charges: 0
* Hospital tier: 1
* City tier: 1
* State ID: 2

1. Fill in the missing values of ‘month’ with Sep and ‘year’ with its average rounded to the nearest integer.

We can replace the missing values in month column by Sep

=IF($C2="?","Sep",$C2)

and to find average rounded,

=IF(B2="?",ROUND(AVERAGE($B$2:$B$2345),0),B2)

1. Determine the most frequently occurring values in the ‘smoker’, 'Hospital tier'and 'City tier' columns, and fill in the missing values accordingly.

In the above case, we have found the most frequently used values and replaced with tier – 2.

1. If any 'State ID' values are missing, consider filling them with 'Unknown' or using another appropriate strategy.

We can use =IF(I2="?","Unknown",I2) to replace the missing State ID.

**Data Transformation:**

**1)** Split the ‘names’ column in the “Customer Names” Table into 3 meaningful columns:

‘Title’, ‘First Name’, and ‘Last Name’.

We can go with the option Text to Columns in DATA tab

**2)** Convert the "NumberOfMajorSurgeries" column in the “Medical Examinations” Table to numerical data by replacing non-numeric characters with meaningful numerical values.

We can use Find and Replace option to change non-numeric to numeric.

We have changed the 1074 non numeric values to numeric value(0) to the given data.

**3)** Check for inconsistencies in the 'Heart Issues' and 'smoker' columns and propose

corrective actions if necessary.

We can change the proper format by using FIND and Replace. 926 replacements done in Heart Issues and 488 replacements done in Smoker rows.

**4)** Create a new column named “Weight Status” that categorizes BMI into different

categories as below:

We can use =IF(B2<18.5,"Underweight",IF(B2<24.9,"Normal Weight",IF(B2<29.9,"Over Weight","Obesity")))

**5)** Create a new column named “Diabetes Status” and fill it as per the information given

below:

The below mentioned format will helps to find the Diabetes Status.

=IF(C2<5.7,"Normal",IF(C2<6.4,"Pre diabetes","Diabetes"))

**6)** Merge ‘year’, ‘month’ and ‘date’ columns in the “Hospitalization Details” Table into one column named ‘Date of Birth’ and format it in ‘DD-MMM-YYYY’ custom format.

We can use =CONCATENATE(O3,"-",N3,"-",M3)

**7)** Calculate the ‘Age’ of each customer based on their ‘Date of Birth’ and the date of

collection of the dataset, which is 8thJune 2023. (Hint: Use the *DATEDIF* function)

We can find the Age, by using =DATEDIF(U3,V3,"Y")

**8)** Format ‘charges’ column as currency ($).

Format changed to currency.

**Data Exploration:**

1. **Customer Names Table:**

* Are there any duplicate Customer IDs in the dataset? If yes, how many?

We have 5 Duplicate customer ID found from the data.

* How many customers are included in the dataset?

There are 2335 Customers included in the dataset.

2. **Medical Examination Table:**

* How many customers have a history of cancer?

We can get the details from the table by using filter, There are 391 customers have a history of cancer.

* How many obese customers have heart issues?

926 customers have heart issues .

* What is the total number of major surgeries performed on customers?

Total 22 number of 3 Major surgeries, 274 number of 2 major surgeries and 965 1 major surgeries performed on customers.

* Calculate the percentage of customers who have undergone any transplants.

6.7% of the customers have undergone transplants.

* Find the average HBA1C value of customers who are smokers.

The average HBA1C value is 6.578.

3. **Hospitalization details Table:**

* Calculate all the Summary statistics for the ‘charges’ column.

We can get the details by using the sum function in charges column.

i.e. Rs. **₹ 3,17,68,896.02**

* Find the average hospitalization charges for customers who are more than 50 years old.

We have =ROUND(AVERAGEIF([Age],">50",[Charges])) function, the average became Rs.**17865.45**

* Compare the total charges across different hospital tiers.

| **Hospital Tier** | **Total Charges** |
| --- | --- |
| Tier 1 | $ 93,10,917.49 |
| Tier 2 | $1,58,99,488.89 |
| Tier 3 | $ 65,58,489.64 |

* Calculate the average charges for people who have more than 2 children.

We can use the =ROUND(AVERAGEIF(E2:E2344,">2",F2:F2344),2)format and we get Rs. **14,217.52** as average charges

* Find the integer average number of children of customers who are less than 40 years old.

We can use =ROUND(AVERAGEIF(O2:O2344,"<40",(E2:E2344)),0) to get the average number of children of customer who are less than 40 yrs. I.e 1.

**Data Analysis:**

* Create a new sheet named “Healthcare", combine all three tables into one, using

Customer ID as the common column, utilizing VLOOKUP.

We can use the table by using VLOOKUP as

=VLOOKUP('Customer Names'!A2,Table7[[Customer ID]:[First Name]],1,)

* Retain the following necessary columns: Customer ID, First Name, BMI, HBA1C, Heart Issues, Any Transplants, Cancer history,numberOfMajorSurgeries, smoker, WeightStatus, Diabetes Status, Date of Birth, charges, Hospital tier, City tier, State ID, Age.

The necessary column obtained by using vlookup in Healthcare.