**Final Project SI 564**

**Name: Preksha Shah**

**Uniqname: prekshah**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

From: Lawrence Summerset

To: DBA team

Hello DBA team!  I hope you are all doing well! We are planning on investing in a hospital setting and beginning our journey in the healthcare industry. But before that, we would like to analyze the place and the services they offer. We have a small sample database that they have provided, which will help us understand their current workflow and profitability.

I have some questions regarding the same and it would be great if you could answer them by the end of this week. Also, it seems that the person who built this database did not leave any documentation, so can you attach an ERD and document it as well?

**Demographics**

1. I would like to know the religious structure of the population we are dealing with. Can you give me the list of any 5 random religions?

**Profitability**

1. The profitability of a hospital depends on the kind of insurance they have in its Payer mix. I would like to see the proportion of deaths among each insurance type, to determine whether the hospital is biased in their treatment and care, considering Private insurance pays more compared to Medicare/Medicaid.
2. Can we determine the average length of stay for Elective and Emergency admissions? The more the length of stay, the more resources are getting used up.

**Public health:**

1. What are the 5 most common types of cancer in the White population in this dataset and can you count the number of diagnosed cases for each type of cancer?
2. What is the survival rate for newborns? This will help us understand whether we want to invest in their pediatric subsidiary or not.

**Hospital Outcomes:**

1. Is there a correlation between the length of admission and death during the hospital stay?
2. Is there a significance of marital status when a patient is diagnosed with breast cancer?

Thank you for your help on this matter. Also, can you asses what was the rationale behind them giving us this dataset and making these design choices?

Have a great week ahead.

Dr. Lawrence Summerset

Head of Research

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

From: DBA team

To: Lawrence Summerset

This is a great opportunity for the company and a great space to work in. I would gladly answer the questions and provide the relevant statistics. I have attached an appendix at the bottom, with the required data tables, documentation, and design concepts.

The answers are as follows:

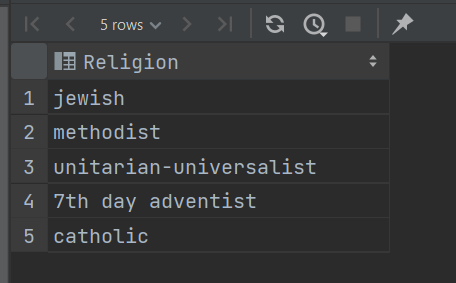
**Demographics**

1. I would like to know the religious structure of the population we are dealing with. Can you give me the list of any 5 random religions?

Query:

select distinct Religion\_Name as "Religion"  
from Hospital\_Admissions\_Data  
join Religion R on Hospital\_Admissions\_Data.Religion\_Type = R.Religion\_Id order by *rand*() limit 5;

Output:



Answer:

These are random 5 religions in this dataset.

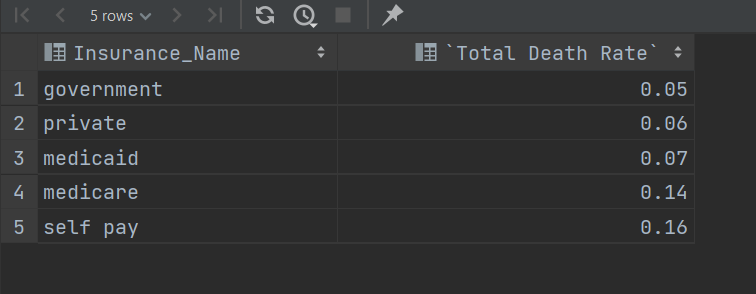
**Profitability**

1. The profitability of a hospital depends on the kind of insurance they have in its Payer mix. I would like to see the proportion of deaths among each insurance type, to determine whether the hospital is biased in their treatment and care, considering Private insurance pays more compared to Medicare/Medicaid.

Query:

select Insurance\_Name, *round*(*count*(Outcome\_Id)/*count*(Insurance\_Type), 2) AS "Total Death Rate"  
from Hospital\_Admissions\_Data had  
join Insurance I on I.Insurance\_Id = had.Insurance\_Type  
left join (SELECT Outcome\_Id from Outcome O where Outcome like "%Dead%") OT on OT.Outcome\_Id = had.Outcome\_Status  
group by Insurance\_Name  
order by *count*(Outcome\_Id)/*count*(Insurance\_Type);

Output:



Answer:

There seems to be a bias in the treatment of those who pay via Medicare and Medicaid, compared to those who pay via Govt/private insurance. This is an area to look at because such a scenario can bring in legal issues and cause loss to us.

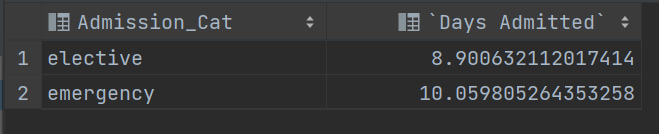
Also, to note, the reimbursement rate is more for Govt/private insurance compared to those of Medicare/Medicaid. As commercial/ private insurance payers pay more/ generate more revenue when compared to payments from Medicare/Medicaid (Lopez et al., 2020)

1. Can we determine the average length of stay for Elective and Emergency admissions?

Query:

select Admission\_Cat, *avg*(Admission\_Length\_Days) as "Days Admitted"  
from Hospital\_Admissions\_Data  
join Admission\_Category AC on AC.Admission\_Cat\_Id = Hospital\_Admissions\_Data.Admission\_Type  
where Admission\_Cat not like "%newborn%" and Admission\_Cat not like "%urgent%"  
group by Admission\_Cat;

Output:



Answer:

As we can see, the emergency department has more average length of stay. The more the length of stay, the more resources are getting used up. Hence, we need to optimize emergency care and invest in resources for that department to avoid loss.

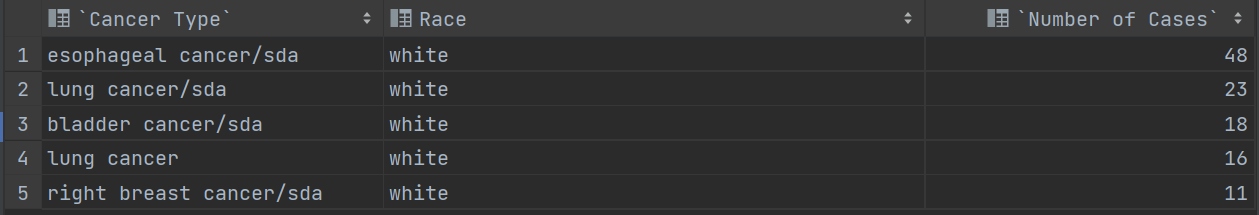
**Public health:**

1. What are the 5 most common types of cancer in the White population in this dataset and can you count the number of diagnosed cases for each type of cancer?

Query:

select Diagnosis as "Cancer Type", R.Race, *count*(Id) as "Number of Cases" from Hospital\_Admissions\_Data  
JOIN Race R on R.Race\_Id = Hospital\_Admissions\_Data.Race  
where Diagnosis like "%cancer%" and R.Race like "%White%"  
group by Diagnosis, R.Race  
order by *count*(Id) desc  
limit 5;

Output:



Answer:

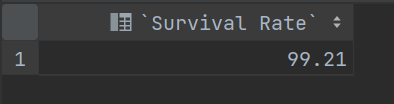
Esophageal cancer, Lung cancer/sda, Bladder Cancer, Lung cancer, and breast cancer are among its top five. We can also recommend they streamline their diagnosis terminology to avoid misinformation.

1. What is the survival rate for newborns? This will help us understand whether we want to invest in their pediatric subsidiary or not.

Query:

select  
 *round*((select *count*(Id) from Hospital\_Admissions\_Data  
 join Admission\_Category AC on Hospital\_Admissions\_Data.Admission\_Type = AC.Admission\_Cat\_Id  
 join Outcome O on O.Outcome\_Id = Hospital\_Admissions\_Data.Outcome\_Status  
 where Admission\_Cat = "newborn" and Outcome = "Alive")/*count*(Id)\*100, 2) as "Survival Rate" FROM Hospital\_Admissions\_Data  
JOIN Admission\_Category A on A.Admission\_Cat\_Id = Hospital\_Admissions\_Data.Admission\_Type  
WHERE Admission\_Cat = "newborn";

Output:



Answer:

The survival rate is 99.21 %. This is a great thing and shows that we should invest in their pediatric subsidiary.

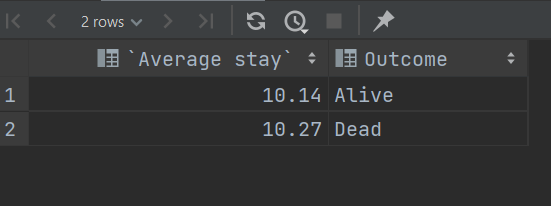
**Hospital Outcomes:**

1. Is there a correlation between the length of admission and death during the hospital stay?

Query:

select *round*(*avg*(Admission\_Length\_Days), 2) as "Average stay", Outcome  
from Hospital\_Admissions\_Data  
join Outcome O on O.Outcome\_Id = Hospital\_Admissions\_Data.Outcome\_Status  
group by Outcome;

Output:



Answer:

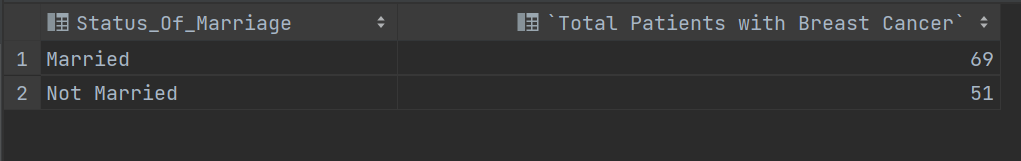
There seems to be no correlation between the two. The average stay of patients is the same for either Outcome.

1. Is there a significance of marital status when a patient is diagnosed with breast cancer?

Query:

select Status\_Of\_Marriage, *count*(Status\_Of\_Marriage) as "Total Patients with Breast Cancer" from Hospital\_Admissions\_Data  
join Marital\_Status MS on MS.Marital\_Status\_Id = Hospital\_Admissions\_Data.Marital\_Status  
where Diagnosis like "%Breast%"  
group by Status\_Of\_Marriage  
order by *count*(Status\_Of\_Marriage) desc;

Output:



Answer:

Though there is not a drastic relation between Breast Cancer and marital status. This area should be explored further and can be used for research purposes.

Hope this helps and in case of any further questions do let me know. Keep me updated on whether we go through with this final project or not. I am excited about it.

Have a great weekend ahead.

Best,

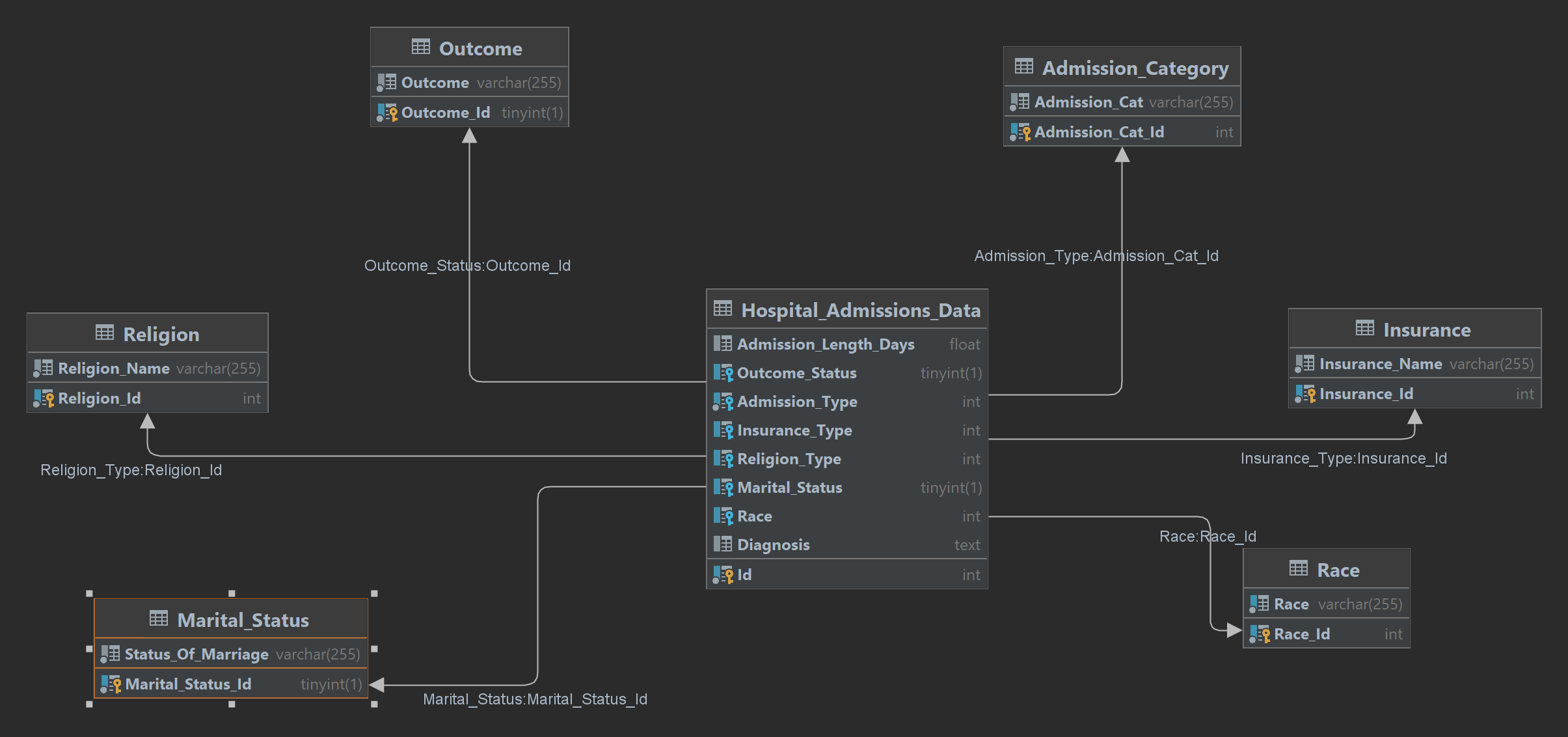
Preksha Shah

DBA Team

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

Appendix:

ERD Diagram:



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

**Documentation and Database design choices:**

The dataset was taken from one of my course work, where we come across Hospital Admissions Dataset. The raw data has 10 attributes, which are then lowered to 9 after cleaning the dataset. There are 58863 rows, each belonging to a unique patient id. There were various field types like categorical data, Boolean, integer, text, and varchar.

Source: <https://drive.google.com/file/d/10DQtxiyIkvFpRADHpH3M0iOb3bAz3Y1J/view?usp=share_link>

I cleaned the raw data after importing it into DataGrip.

For that:

1. Checked for any duplicate or repeated entries , or loss of information for the same id in the table made from raw data.

select ID, *COUNT*(1) from HospitalAdmissionsData\_1 group by ID ORDER BY ID DESC ;

I de-identified that data from the raw data. This was done to ensure patient privacy as per HIPAA rules.

1. De-identified data

ALTER TABLE HospitalAdmissionsData  
DROP COLUMN Hospital\_id;

1. Deleted unnecessary data

ALTER TABLE HospitalAdmissionsData  
DROP COLUMN EnglishLanguage\_1;

1. Remove negative values from Admission\_Length\_Days

update Hospital\_Admissions\_Data h set Admission\_Length\_Days = NULL where Admission\_Length\_Days < 0;

**Data Normalization:**

To reduce repetition and redundancy, I created sub-tables from it as follows:

1. **Hospital\_Admissions\_Data**

Id: Autoincrement, int, PrimaryKey, not null

Admission\_Length\_Days: No. Of Days in Hospital, float

Outcome\_Status: Outcome of Treatment (Boolean), 0 is alive, 1 is dead.

Admission\_Type: Category ID under which admitted, int

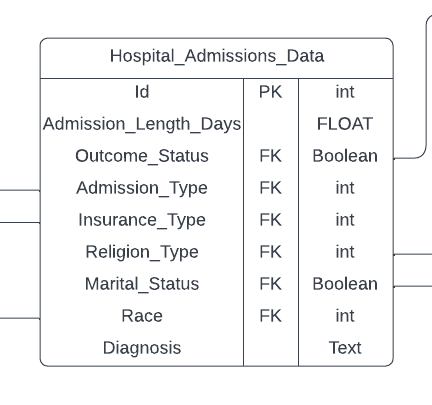
Insurance\_Type: Category ID for type of Health insurance, int

Religion\_Type: Category ID of Religion, int

Marrital\_Status: Status of marriage(Boolean), 0 is not married, 1 is married

Race: Category ID of Race, int

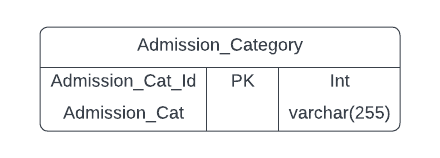
Dx: Diagnosis (text string)



1. **Admission\_Category**

Admission\_Cat\_Id : Category Id, int, autoincrement, not null, PrimaryKey

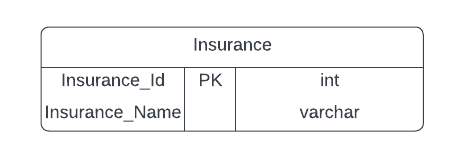
Admission\_Cat : Category Name, varchar(255)



1. **Insurance**

Insurance\_Id : Category Id, int, autoincrement, not null, PrimaryKey

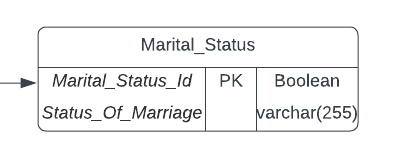
Insurance\_Name: Category Name, varchar(255)



1. **Marital\_Status**

Marital\_Status\_Id : Boolean, Tinyint, notnull, Primarykey

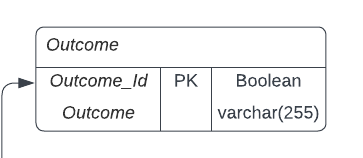
Status\_Of\_Marriage : Married/NotMarried, varchar(255)



1. **Outcome**

Outcome\_Id : Boolean, Tinyint, notnull, Primarykey

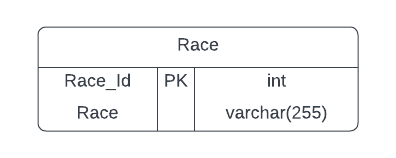
Outcome: Alive/Dead, varchar(255)



1. **Race**

Race\_Id : Category Id, int, autoincrement, not null, PrimaryKey

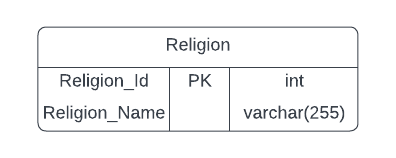
Race: Category Name, varchar(255)



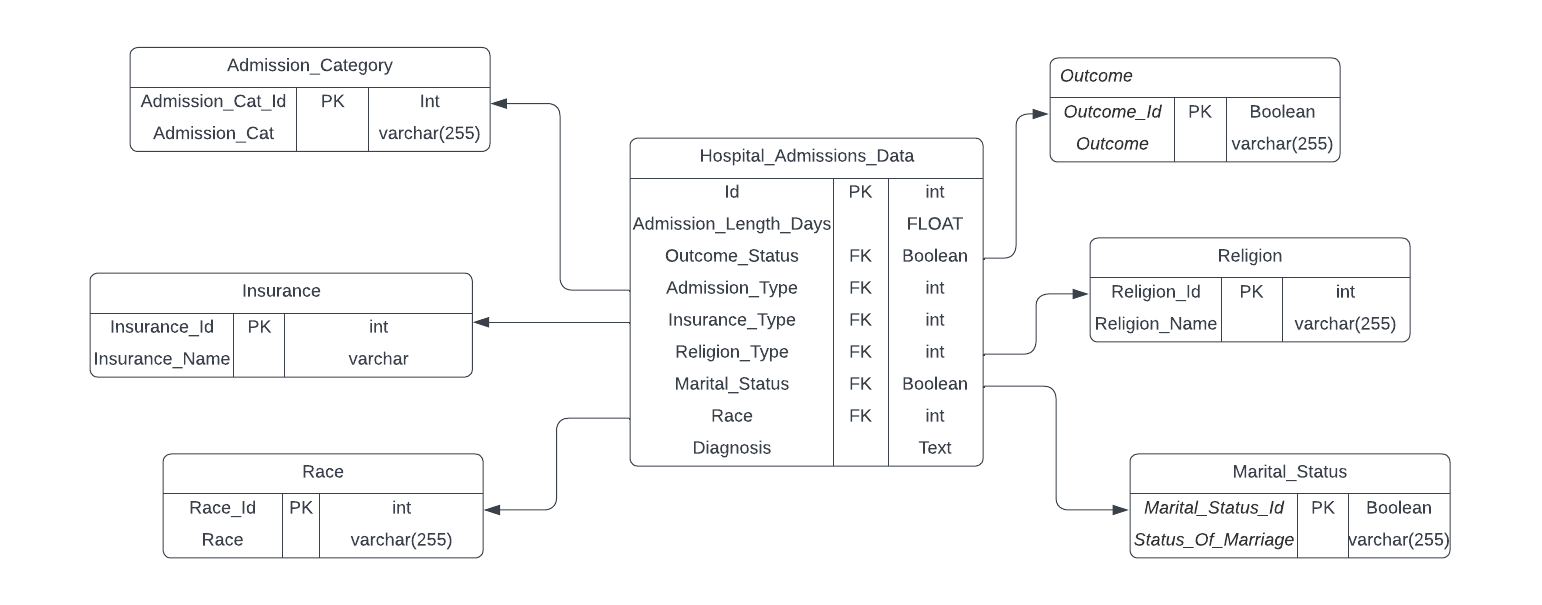
1. **Religion**

Religion\_Id : Category Id, int, autoincrement, not null, PrimaryKey

Religion\_Name: Category Name, varchar(255)



ERD Made manually:



The above tables were made from the raw table data, and then main table was created using following queries:

update HospitalAdmissionsData h set Religion\_Type = (select id from Religion\_Type re where re.Religion = h.Religion\_Type)

update Hospital\_Admissions\_Data h set Admission\_Type = (select Id from Admission\_Category ac where ac.Admission\_Cat = h.Admission\_Type)

update Hospital\_Admissions\_Data h set Insurance\_Type = (select Insurance\_Id from Insurance i where i.Insurance\_Name = h.Insurance\_Type)

update Hospital\_Admissions\_Data h set Race = (select Race\_Id from Race r where r.Race = h.Race)

Further, I also made foreign keys and joined the tables accordingly.

I have attached three MySQL dump files, so that it covers all the changes I made in the dataset throughout the week at different intervals.

The most latest one is Done\_Finally\_With\_Project

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