Analysis Questions

1. To create a summary of drugs used by ethnicity, you first need to define a counting method for drugs. I chose to go with prescription counts rather than the amount of drug prescribed since many drugs have different units and this comparison would not necessarily be the most informative. Using a SQL query, we can obtain the prescription count data for every ethnicity drug pair, and then find the top for each group by dropping excess data in post.

```
Most perscribed drugs by ethnicity:
AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNIZED TRIBE: 5% Dextrose, 54
ASIAN: D5W, 27
BLACK/AFRICAN AMERICAN: Insulin, 60
HISPANIC OR LATINO: 5% Dextrose, 28
HISPANIC/LATINO - PUERTO RICAN: 0.9% Sodium Chloride, 1290
OTHER: NS, 11
UNABLE TO OBTAIN: 0.9% Sodium Chloride, 28
UNKNOWN/NOT SPECIFIED: D5W, 41
WHITE: Potassium Chloride, 508
```

2. The main query behind this question is finding all procedures performed and the age at which the patient was for the procedure. The age comes from the admission date minus the patient's date of birth. Using a custom function, we can add an age group column to the data using pandas, which is easier to do than in SQL itself. Finally, the top outputs are selected for print to see the results. It is clear that Venous cath NEC is the most common procedure across age groups, which makes sense given that most patients have some sort of medicine administered this way.

```
<=19:
Venous cath NEC: 2
Temporary tracheostomy: 1
Packed cell transfusion: 1
Vertebral fx repair: 1
Percu endosc gastrostomy: 1
Skin closure NEC: 1
Closed bronchial biopsy: 1
Open reduc-int fix femur: 1
Entral infus nutrit sub: 1
Fus/refus 2-3 vertebrae: 1
Spinal tap: 1
Applic ext fix dev-femur: 1
Remove imp device-femur: 1
```

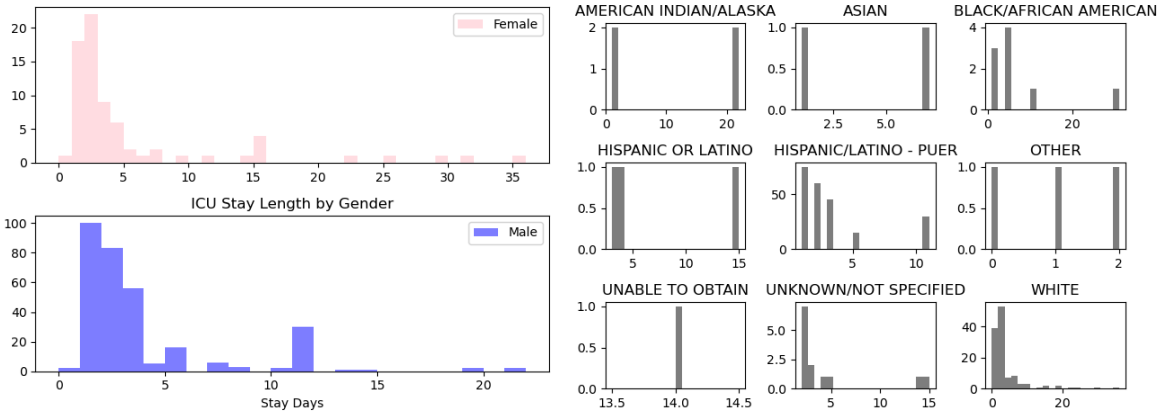
```
Incision of lung: 1
Atlas-axis fusion: 1
Ot cerv fusion post/post: 1
Cl fx reduc-femur: 1
Interruption vena cava: 1
Other skeletal traction: 1

20-49:
Venous cath NEC: 9
Entral infus nutrit sub: 7
Insert endotracheal tube: 6

50-79:
Venous cath NEC: 26
Entral infus nutrit sub: 22
Packed cell transfusion: 13

>=80:
Venous cath NEC: 19
Packed cell transfusion: 13
Insert endotracheal tube: 8
```

3. A SQL query to obtain all of the hospital stays, their length, and the gender and ethnicity of the subject is the backbone of this question. From there, the data is filtered for plotting by both gender and ethnicity separately. The plots below are histograms for stay lengths in days, followed by a dataframe which shows the number of patients who had multiple stays in the dataset. This information is important in case we find that patients with multiple stays have a separate distribution of stay lengths compared to the average patient being admitted due to being more likely to be unhealthy.



Cassandra Tables

The Cassandra tables are generated using the SQL queries and data from the analysis questions under the constraints of no JOIN statements in Cassandra and no aggregations in the tables. The data is saved in a .csv file to be validated against the .csv files from the original analysis questions.