Findings & Justifications

I created a Structured Query Language (SQL) files of all the Dataset provided in the question.

1. HospitalProfiling.sql was created.
2. Cleaned Data of HospitalProfiling.sql by taking the avg Empoyee count and stored in HospitalProfiling\_con\_avg.sql

Few Columns were added (Year\_Average,Year\_Median,Max\_Year,Min\_Year,First\_Quarter\_Sum,Second\_Quarter\_Sum,Third\_Quarter\_Sum,Fourth\_Quarter\_Sum,First\_Quarter\_Avg, Second\_Quarter\_Avg,Third\_Quarter\_Avg,Fourth\_Quarter\_Avg,First\_Half\_Year\_Total,First\_Half\_Year\_Average,Second\_Half\_Year\_Total,Second\_Half\_Year\_Average) to HospitalRevenue.csv to make HospitalRevenue\_added\_cols.sql

1. Removed "NO District Available" District\_Dd rows from HospitalRevenue\_added\_cols.sql
2. Added Buy\_or\_not column to ProjectedRevenue.sql
3. Removed "NO District Available" District\_id rows from ProjectedRevenue.sql
4. Created table `hospitalprofiling+revenue.sql` which contains the join of Hospital Profile and Hospital Revenue to get the Hospital\_Employees with the Revenues table.
5. Checked for duplicates in hospitalprofiling+revenue table, none found :)
6. Created `projectedrevenue\_added\_cols.sql` by adding Buy\_or\_not as 1
7. Created `hospitalprofiling\_projectedrevenue.sql` by joining `projectedrevenue\_added\_cols` & `hospitalprofiling+revenue` tables
8. Created `hospitalprofiling\_minus\_revenue.sql` by subtracting `hospitalprofiling\_projectedrevenue` from `hospitalprofiling+revenue` tables
9. Created dataset.csv by combining 2 CSV files `hospitalprofiling\_projectedrevenue` & `hospitalprofiling\_minus\_revenue`

\*\*\*Assuming the `hospitalprofiling\_minus\_revenue.csv` has the not sold data as they were not found to have the predicted values when we created `hospitalprofiling\_projectedrevenue`\*\*\*

This dataset.csv was my Main Dataset.

Revenue Prediction:

1. I first chose Regression to find predicted Revenues.
2. Divided this dataset into 70%, 30% to create the training and testing dataset respectively.
3. Drew the correlation plots, box plots and summary plots to have a quick look at the trainset.
4. Next ran the regression algorithm using the `glm` library available in R.
5. Tested this against the testing set. Got fairly decent results.
6. Then scored the Solution.csv file with the predicted value from my model.

Buy\_or\_not Prediction:

1. I then chose the ADA Boost algorithm with 50 iterations to predict this target variable.
2. The same 70, 30 ratio was used for training and testing datasets.
3. After the testing gave 100% accuracy I scored the Solution.csv file with my new model.

The final Solution.csv that was generated was uploaded in the SQL table.

Wherever the Buy\_or\_not attribute was 0, the Predicted revenue value was made 0.

Queries used:

--- To find Duplicate data:

SELECT \*,count(\*) FROM `mytable` GROUP BY Hospital\_ID,Region\_ID,District\_ID,Instrument\_ID,Month\_1,Month\_2,Month\_3,Month\_4,Month\_5,Month\_6,Month\_7,Month\_8,Month\_9,Month\_10,Month\_11,Month\_12,Year\_Total,Year\_Average,Year\_Median,Max\_Year,Min\_Year,First\_Quarter\_Sum,Second\_Quarter\_Sum,Third\_Quarter\_Sum,Fourth\_Quarter\_Sum,First\_Quarter\_Avg,Second\_Quarter\_Avg,Third\_Quarter\_Avg,Fourth\_Quarter\_Avg,First\_Half\_Year\_Total,First\_Half\_Year\_Average,Second\_Half\_Year\_Total,Second\_Half\_Year\_Average HAVING count(\*)>1 ORDER BY `Month\_1` ASC

--- Hospital Profile and Hospital Revenue join

select hr.Hospital\_ID,hr.Region\_ID,hr.District\_ID,hr.Instrument\_ID,Month\_1,Month\_2,Month\_3,Month\_4,Month\_5,Month\_6,Month\_7,Month\_8,Month\_9,Month\_10,Month\_11,Month\_12,Year\_Total,Year\_Average,Year\_Median,Max\_Year,Min\_Year,First\_Quarter\_Sum,Second\_Quarter\_Sum,Third\_Quarter\_Sum,Fourth\_Quarter\_Sum,First\_Quarter\_Avg,Second\_Quarter\_Avg,Third\_Quarter\_Avg,Fourth\_Quarter\_Avg,First\_Half\_Year\_Total,First\_Half\_Year\_Average,Second\_Half\_Year\_Total,Second\_Half\_Year\_Average,hp.Hospital\_employees from hospitalprofiling\_conv\_avg hp,hospitalrevenue\_added\_cols hr where hr.Hospital\_ID=hp.Hospital\_ID and hr.District\_ID=hp.District\_ID

--- Hospital Revenue and Projected Revenue join

select hpr.Hospital\_ID,hpr.Region\_ID,hpr.District\_ID,hpr.Instrument\_ID,Month\_1,Month\_2,Month\_3,Month\_4,Month\_5,Month\_6,Month\_7,Month\_8,Month\_9,Month\_10,Month\_11,Month\_12,Year\_Total,Year\_Average,Year\_Median,Max\_Year,Min\_Year,First\_Quarter\_Sum,Second\_Quarter\_Sum,Third\_Quarter\_Sum,Fourth\_Quarter\_Sum,First\_Quarter\_Avg,Second\_Quarter\_Avg,Third\_Quarter\_Avg,Fourth\_Quarter\_Avg,First\_Half\_Year\_Total,First\_Half\_Year\_Average,Second\_Half\_Year\_Total,Second\_Half\_Year\_Average,Hospital\_employees,Annual\_Projected\_Revenue,Buy\_or\_not from `hospitalprofiling+revenue` hpr,projectedrevenue\_added\_cols pr where hpr.Hospital\_ID=pr.Hospital\_ID and hpr.District\_ID=pr.District\_ID and hpr.Instrument\_ID=pr.Instrument\_ID

\*\*\* `projectedrevenue\_added\_cols` - `hospitalprofiling\_projectedrevenue` [Sold instruments, but not in my inventory

select Hospital\_ID,District\_ID,Instrument\_ID from projectedrevenue\_added\_cols pra where not exists (select Hospital\_ID,District\_ID,Instrument\_ID from hospitalprofiling\_projectedrevenue hppr where pra.Hospital\_ID=hppr.Hospital\_ID and pra.District\_ID=hppr.District\_ID and pra.Instrument\_ID=hppr.Instrument\_ID)

--- `hospitalprofiling+revenue` - `projectedrevenue\_added\_cols` [NOT Sold instruments, but in my inventory]

select pra.Hospital\_ID,pra.District\_ID,pra.Instrument\_ID,Month\_1,Month\_2,Month\_3,Month\_4,Month\_5,Month\_6,Month\_7,Month\_8,Month\_9,Month\_10,Month\_11,Month\_12,Year\_Total,Year\_Average,Year\_Median,Max\_Year,Min\_Year,First\_Quarter\_Sum,Second\_Quarter\_Sum,Third\_Quarter\_Sum,Fourth\_QuarterSum,First\_Quarter\_Avg,Second\_Quarter\_Avg,Third\_Quarter\_Avg,Fourth\_Quarter\_Avg,First\_Half\_Year\_Total,First\_Half\_Year\_Average,Second\_Half\_Year\_Total,Second\_Half\_Year\_Average,Hospital\_employees from `hospitalprofiling+revenue` pra where not exists (select Hospital\_ID,District\_ID,Instrument\_ID from projectedrevenue\_added\_cols hppr where pra.Hospital\_ID=hppr.Hospital\_ID and pra.District\_ID=hppr.District\_ID and pra.Instrument\_ID=hppr.Instrument\_ID)

Also,

Based on Correlation plots, removed Min \_Year & Hospital\_employees Column [Terribly low correlation]; removed First.Quarter.Sum,Second.Quarter.Sum,Third.Quarter.Sum,Fourth.Quarter.Sum,Second.Half.Year.Total,Year.Total Columns since same correlation as their Average counterparts hence, created ds1.csv

Normalized ds1 by recentering.

--Abhishek Karan