Flatiron Health - Interview Exercise

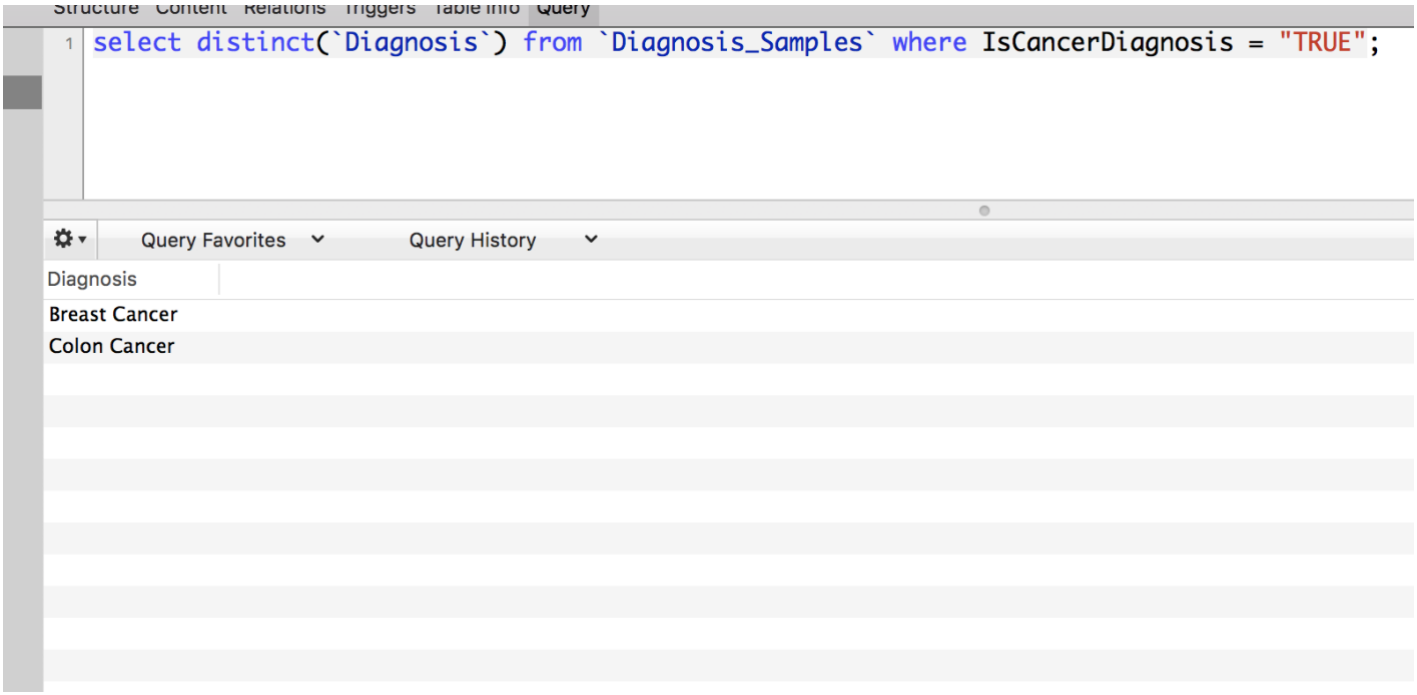
By Harish Puvvada

Language: ***MySQL***
Application: ***Sequel Pro***

1a) Which types of cancer does the clinic see patients for?

Query: *select distinct(`Diagnosis`) from `Diagnosis_Samples` where IsCancerDiagnosis = "TRUE";*

Results:



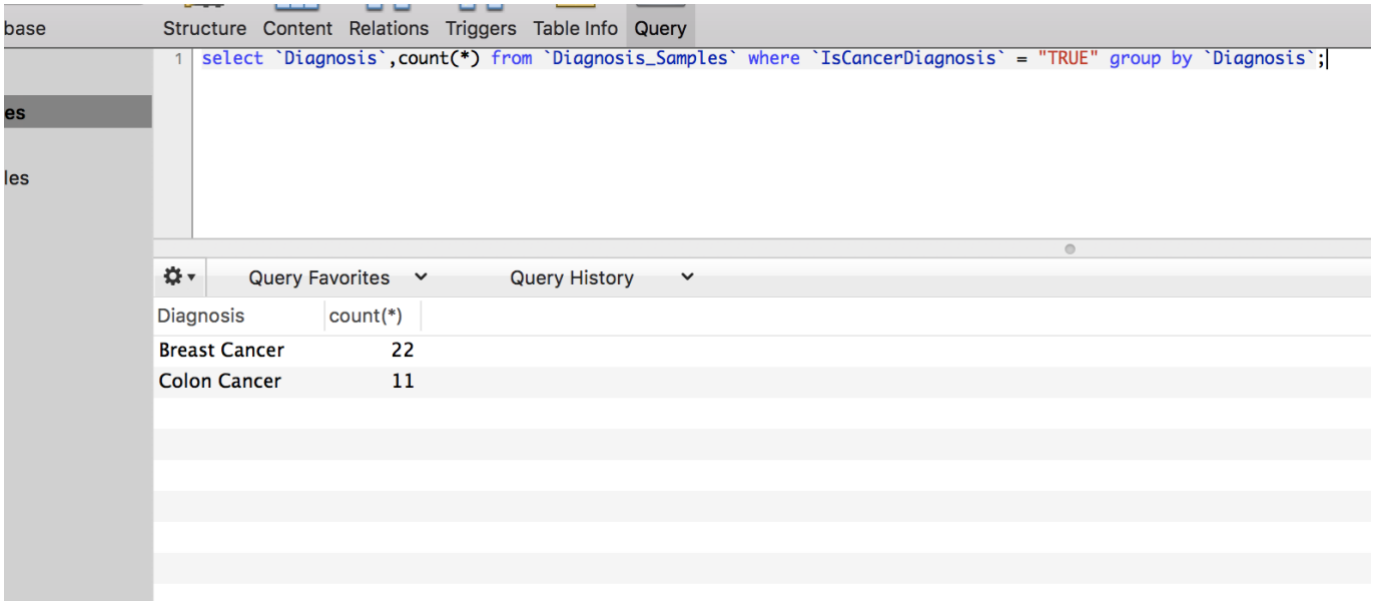
Query Explanation: I used **distinct** to find different types of cancer.

But there were patients who came for diagnosis but did not have cancer. To filter out those patients, I used **IsCancerDiagnosis flag.**

1b) How many patients does the clinic have for each cancer type?

Query: *select `Diagnosis`,count(*) from `Diagnosis_Samples` where `IsCancerDiagnosis` = "TRUE" group by `Diagnosis`;*

Results:



Query Explanation: I used **IsCancerDiagnosis flag** to filter out patients with cancer and grouped by **Diagnosis.**

2a) How long after being diagnosed do patients start treatment for each cancer type?

Query: *select DT.`PatientID`,Datediff(DT.min_Trt,DT.min_Diag) from*
(select TS.`PatientID`,
min(STR_TO_DATE(TS.`TreatmentDate`,'%m/%d/%Y')) as min_Trt,
min(STR_TO_DATE(DS.`DiagnosisDate`,'%m/%d/%Y')) as min_Diag
from `Diagnosis_Samples` DS inner join `Treatment_Samples` TS
on DS.`PatientID`=TS.`PatientID`
where TS.`TreatmentDate`> DS.`DiagnosisDate` group by DS.`PatientID`) DT;

Results:

```
select DT.`PatientID`,Datediff(DT.min_Trt,DT.min_Diag) from
(select TS.`PatientID`,
min(STR_TO_DATE(TS.`TreatmentDate`,'%m/%d/%Y')) as min_Trt,
min(STR_TO_DATE(DS.`DiagnosisDate`,'%m/%d/%Y')) as min_Diag
from `Diagnosis_Samples` DS inner join `Treatment_Samples` TS
on DS.`PatientID`=TS.`PatientID`
where TS.`TreatmentDate`> DS.`DiagnosisDate` group by DS.`PatientID`) DT;
```

Query Favorites Query History

PatientID	Datediff(DT.min_Trt,DT.min_Diag)
2038	3
2120	23
2175	4
2407	6
2425	4
2462	25
2763	4
2770	6
3095	3
3757	11
3948	4
4256	25
4354	5
4374	5
4692	3
5259	4
6281	4
6837	5
6877	7
6889	8
6922	21
7230	7
7242	6
7796	5
7976	30
9331	6

(explanation in next page)

Query Explanation: I converted the varchar dates to date format and then found out the earliest **DiagnosisDate** and **TreatmentDate** and then found difference between them to find out the time taken by patient to get the treatment after diagnosis.

Note: Here, if a patient has more than one cancer, then we cannot find out the time taken for the other cancers except the first one.

Data Insufficiency: There is no treatment ID for each patient. So, if a patient has more than one cancer We need a column like **treatment ID** in which can differentiate between different rows in **Treatment_Samples** table (*this should be a foreign key in **Diagnosis_Samples** table*).

2b) Are there patients which are not treated at the practice?

Query:

SELECT DS.`PatientID` FROM `Diagnosis_Samples` DS left join `Treatment_Samples` TS on DS.`PatientID`=TS.`PatientID` where TS.`PatientID` is NULL;

Results:

```
SELECT DS.`PatientID` FROM `Diagnosis_Samples` DS
left join `Treatment_Samples` TS
on DS.`PatientID`=TS.`PatientID`
where TS.`PatientID` is NULL;
```

Query Favorites Query History

PatientID
2634
5657
6840
7937
8615
8827

Explanation:

I am using a **Left join** which will include all the patient ids in **Diagnosis_Samples** table. Then I filtered out rows with **PatientId** (of Treatment_Samples table) with NULL value

Observation: Here one common thing I could see between these patients was that they are all patients who were not having cancer. Although there are more patients who did not have cancer but treated.

3) After being treated with a first line of treatment (a drug or combination of drugs), what proportion of patients go on to be treated with a second line of treatment?

From the given data, it was not clear which treatment dates are for second line of treatment.

Assumption: A & B drugs - chemotherapy as First line of treatment
C drug - immunotherapy as Second line of treatment

Query to find patients who go for second line of treatment:

select distinct(`PatientID`) from `Treatment_Samples` where `DrugCode` = "C";

Query to know the proportion of patients who go for second line of treatment:

```
select (
(select count(*) from (select distinct(`PatientID`) from `Treatment_Samples` where `DrugCode` = "C")TS)/
(select count(distinct(DS.`PatientID`)) from `Diagnosis_Samples` DS inner join `Treatment_Samples` TS on
DS.`PatientID`=TS.`PatientID` where DS.`IsCancerDiagnosis`="TRUE")
) as Proportion;
```

Explanation: I found the number of patients who used drug C and divided it by total number of patients who had cancer by using inner join between **Diagnosis_Samples** & **Treatment_Samples** table and **IsCancerDiagnosis** flag**.**

Results:

```
select (
(select count(*) from (select distinct(`PatientID`) from `Treatment_Samples` where `DrugCode` = "C")TS)/
(select count(distinct(DS.`PatientID`)) from `Diagnosis_Samples` DS inner join `Treatment_Samples` TS on
DS.`PatientID`=TS.`PatientID` where DS.`IsCancerDiagnosis`="TRUE")
) as Proportion;
```

Query Favorites Query History

Proportion
0.6071

Alternative Assumptions:

Second line of treatment can be after a long time after first line of treatment or it could be immediately after first line of because of side effects or any other reason.

4) How do the drugs used at the clinic compare in terms of duration of therapy?

Query: *select `PatientID`,`DrugCode`,*
DATEDIFF(max(STR_TO_DATE(`TreatmentDate`,'%m/%d/%Y')),
min(STR_TO_DATE(`TreatmentDate`,'%m/%d/%Y'))) as "Duration"
from `Treatment_Samples` group by `PatientID`,`DrugCode`;

Results:

```
select `PatientID`,`DrugCode`,
DATEDIFF(max(STR_TO_DATE(`TreatmentDate`,'%m/%d/%Y')),
         min(STR_TO_DATE(`TreatmentDate`,'%m/%d/%Y'))) as "Duration"
from `Treatment_Samples` group by `PatientID`,`DrugCode` ;
```

Query Favorites Query History

PatientID	DrugCode	Duration
2038	A	27
2038	B	27
2120	A	37
2120	B	37
2175	A	41
2175	B	41
2407	A	45
2407	B	45
2425	A	51
2425	B	51
2462	A	52
2462	B	52
2763	A	60
2763	B	60
2770	A	58
2770	B	58
3095	A	61
3449	C	66
3757	C	87
3948	A	62
3948	B	62
3948	C	85
4256	C	68
4354	C	70
4374	C	73
4692	A	42

Query Explanation: I grouped by **PatientID, DrugCode** and then subtracted earliest date of treatment from recent date of treatment. I used **Str_to_date** to convert given varchar type to date format and then found the difference using **DATEDIFF.**

Note: If a patient used same drug for more than one cancer, this query won't give desired output. This can be resolved the same way as I explained in **2a** problem.

----------Thank You----------