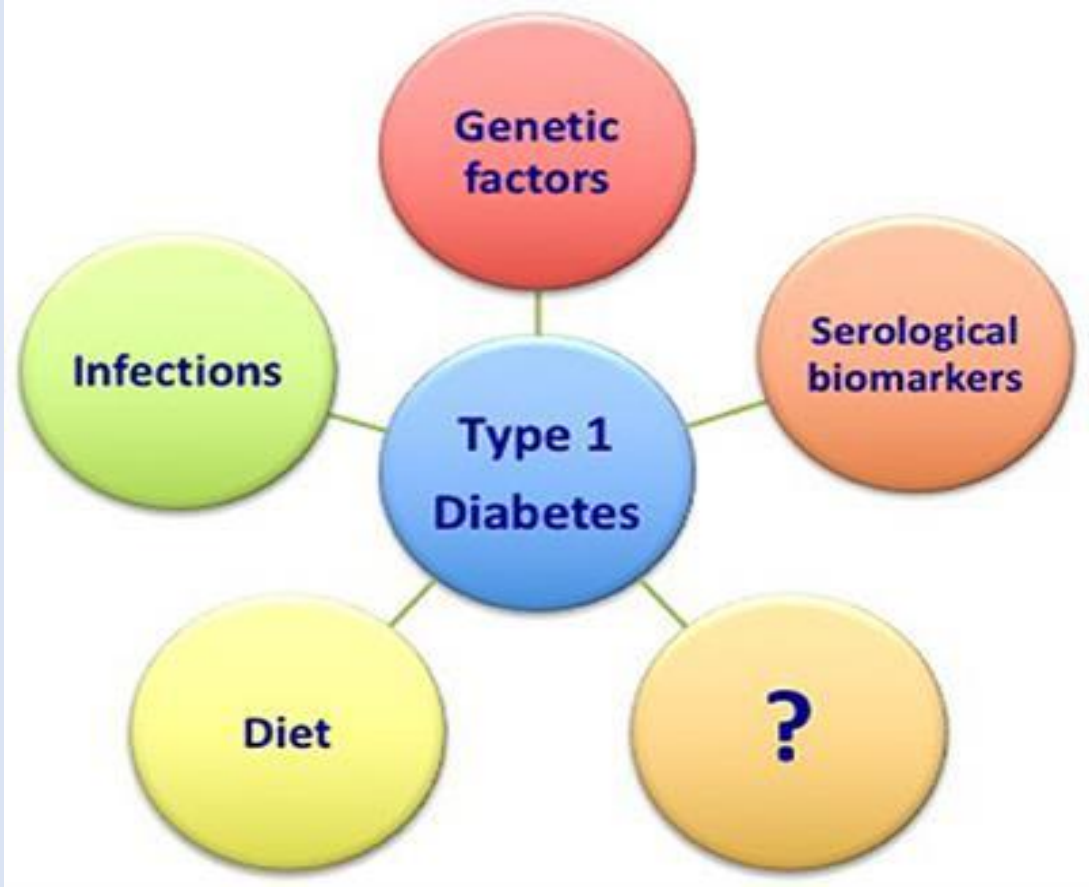
DIABETES PREDICTION ANALYSIS



JESSICA ATARI-ATA

Data Acquisition

- This Analysis was conducted using a comprehensive dataset provided by **PSYLIQ.**

Project Highlights

- A detailed analysis was conducted for Patients with diabetes based on the following factors:
- Age, smoking history, heart disease and so on .

Project Tool

- SQL (BigQuery)

Data Cleaning Process

- Checked for Duplicates, found None.

```sql
-- Checking for Duplicates
SELECT
  EmployeeName,Patient_id,gender,age,hypertension,heart_disease,smoking_history,bmi,HbA1c_level, blood_glucose_level, diabetes
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
GROUP BY
  EmployeeName,Patient_id,gender,age, hypertension,heart_disease,smoking_history,bmi,HbA1c_level, blood_glucose_level,diabetes
HAVING COUNT(*) > 1;
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

There is no data to display.

Checked for Null Values, found none

```sql
-- Checking for Null Values
SELECT
  EmployeeName,Patient_id,gender,age,hypertension,heart_disease,smoking_history,bmi,HbA1c_level, blood_glucose_level, diabetes
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
WHERE EmployeeName IS NULL
OR Patient_id IS NULL
OR gender IS NULL
OR age IS NULL
OR Hypertension IS NULL
OR heart_disease IS NULL
OR smoking_history IS NULL
OR bmi IS NULL
OR HbA1c_level IS NULL
OR blood_glucose_level IS NULL
OR diabetes IS NULL;
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

There is no data to display.

1). Retrieve the Patient_id and ages of all patients.

```sql
-- Retrieve Patient_id and Ages of all patients
SELECT
  EmployeeName,
  Patient_id,
  CAST(age AS INT64) AS Age
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
ORDER BY
  Patient_id,
  Age
```

Query results

JOB INFORMATION | RESULTS | CHART | PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	EmployeeName	Patient_id	Age
1	SHERYL BREGMAN	PT1000	34
2	JOHN HOFFMAN	PT10000	46
3	Meredith H Reddoch-Ho	PT100000	45
4	Minouche Kandel	PT100001	57
5	Mose Thornton	PT100002	72
6	Helen H Chong	PT100003	44
7	Marvin M Mouton	PT100004	62
8	Edward A Ang	PT100005	21
9	Gordon G Leong	PT100006	62
10	Judith Reyes	PT100007	3
11	Tara L Croan	PT100008	48
12	Brian M DeNave	PT100009	67
13	ROBERTO VALLADARES	PT10001	27
14	Jennifer J Pascual	PT100010	41

2. Select all female patients who are older than 40.

```sql
-- Female Patients older than 40
SELECT
  EmployeeName,
  Gender,
  CAST(age AS INT64) AS Age
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
WHERE
  Gender ='Female'
  AND Age >40
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	EmployeeName	Gender	Age
1	FRANCIS HAGAN	Female	42
2	MARIALUZ BANARES	Female	42
3	RICHARD ACERET	Female	42
4	MICHAEL MITCHELL	Female	42
5	AMOD DHAKAL	Female	42
6	RENE QUIAMBAO	Female	42
7	CHARLES SHEEHAN	Female	42
8	DAMIAN DAVIS	Female	42
9	ANNE KOFMAN	Female	42
10	CORY DECKER	Female	42
11	DANIEL IP	Female	42
12	DANIELLE THOMAS	Female	42
13	VIRGILIO RAFANAN	Female	42
14	YVETTE WILLIAMS-DUBRIWNY	Female	42

3). Calculate the average BMI of patients.

```sql
-- Average BMI of Patients
SELECT
  CAST(AVG(bmi) AS INT64) AS Average_BMI
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
```

Query results

JOB INFORMATION | RESULTS | CHART | PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	Average_BMI
1	27

4). List patients in descending order of blood glucose levels.

```sql
-- Blood glucose level of patients in descending order
SELECT
  EmployeeName,
  Patient_id,
  Blood_glucose_level
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
ORDER BY
  Blood_glucose_level
  DESC
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	EmployeeName	Patient_id	Blood_glucose_level
1	Carlos Imbong	PT51740	300
2	Santiago Lagandaon	PT49657	300
3	Yu Qiong Wen	PT63081	300
4	Rita L Kearns	PT76623	300
5	STEPHEN ZOLLMAN	PT5536	300
6	Rosa Narvaez	PT57767	300
7	Teresa Chan	PT45322	300
8	Viet Q Ha	PT78543	300
9	KATHLEEN GULBENGAY	PT36031	300
10	Michael Callejas	PT59205	300
11	Jeff Fisher	PT54591	300
12	Theresa Smith	PT65140	300
13	Sergey Trofimenko	PT89757	300
14	Flor D Roman	PT99809	300

5). Find patients who have hypertension and diabetes.

```sql
-- Patients who have hypertension and diabetes
SELECT
  EmployeeName,
  Hypertension,
  Diabetes
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
WHERE
  Hypertension <> 0
  AND Diabetes <> 0

```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	EmployeeName	Hypertension	Diabetes
1	Richard S Burns	1	1
2	Matelina Alexander	1	1
3	Jeffrey Black	1	1
4	Anthony Branchcomb	1	1
5	Jason Jefferson	1	1
6	Irving McKnight	1	1
7	Joseph B Rocero	1	1
8	Richard Sucre	1	1
9	Justin Morgan	1	1
10	ALISON YEE	1	1
11	Xiao Dong Cai	1	1
12	Jeffrey E Airth	1	1
13	Barbara L Moss	1	1
14	Randy F Acosta	1	1
15	MICHELLE SHINN	1	1

6). Determine the number of patients with heart disease.

➔ The number of patients with heart disease in this dataset is 3,942.

```sql
-- Number of Patients with Heart Disease
SELECT
  COUNT(Heart_Disease) AS Heart_Disease
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
WHERE
  Heart_Disease <> 0
```

Query results

JOB INFORMATION | RESULTS | CHART | PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	Heart_Disease
1	3942

7). Group patients by smoking history and count how many smokers and non-smokers there are.

```sql
--Group patients by smoking history and count how many smokers and non-smokers there are.
select Smoking_history,COUNT(Smoking_history) As Total_Smoking_history
  from `psyliq-intern.Diabetes_prediction.Diabetes  2`
  where Smoking_history is not null
group by Smoking_history
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | E

Row	Smoking_history	Total_Smoking_history
1	never	35095
2	No Info	35816
3	not current	6447
4	current	9286
5	ever	4004
6	former	9352

8) Retrieve the Patient_ids of patients who have a BMI greater than the average BMI.

```sql
-- Patients with greater than average BMIs
-- Average BMI is 27
SELECT
  Patient_id,
  CAST(bmi AS int64) AS BMI
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
Where
  BMI > 27
```

Query results

JOB INFORMATION | RESULTS | CHART | PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	Patient_id	BMI
1	PT50982	57
2	PT4330	57
3	PT22555	88
4	PT41474	49
5	PT13977	51
6	PT68256	48
7	PT97017	51
8	PT59611	52
9	PT90093	49
10	PT96442	51
11	PT72360	48
12	PT5314	50
13	PT40882	51

9). Find the patient with the highest HbA1c level and the patient with the lowest HbA1clevel.

➔ Patients with HbA1c level 9 were the highest and Patients with HbA1c level 3.5 were the lowest as shown below:

```
1  -- Patients with the Highest HbA1c level
2  SELECT Patient_id, MAX(HbA1c_level) AS Highest_HbA1c_level
3  from `psyliq-intern.Diabetes_prediction.Diabetes  2`
4  where patient_id is not null
5  Group by Patient_id
6  Order by Highest_HbA1c_level DESC
7
8
9
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON

Row	Patient_id	Highest_HbA1c_level
1	PT40453	9.0
2	PT44779	9.0
3	PT53598	9.0
4	PT71476	9.0
5	PT73983	9.0
6	PT36071	9.0
7	PT59549	9.0
8	PT11499	9.0
9	PT11719	9.0
10	PT1866	9.0
11	PT87900	9.0
12	PT56655	9.0
13	PT78316	9.0

```
1  -- Patients with the Lowest HbA1c level
2  SELECT Patient_id, MIN(HbA1c_level) AS Lowest_HbA1c_level
3  from `psyliq-intern.Diabetes_prediction.Diabetes  2`
4  where patient_id is not null
5  Group by Patient_id
6  Order by Lowest_HbA1c_level ASC
7
8
9
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON

Row	Patient_id	Lowest_HbA1c_level
1	PT891	3.5
2	PT1141	3.5
3	PT1841	3.5
4	PT1940	3.5
5	PT2032	3.5
6	PT5457	3.5
7	PT6806	3.5
8	PT7552	3.5
9	PT9032	3.5
10	PT9034	3.5
11	PT10302	3.5
12	PT10389	3.5
13	PT11114	3.5

10). Calculate the age of patients in years (assuming the current date as of now).

```
1  -- Calculate the age of patients in years (assuming the current date as of now).
2  SELECT Patient_id,
3  Cast (Age_ AS int64) AS Age
4   FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
5   Order by Age DESC
```

Query results

JOB INFORMATION | RESULTS | CHART | PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	Patient_id	Age
2	PT91186	80
3	PT65525	80
4	PT22551	80
5	PT59430	80
6	PT4009	80
7	PT85234	80
8	PT39672	80
9	PT60419	80
10	PT36000	80
11	PT60494	80
12	PT78666	80
13	PT1066	80
14	PT76319	80
15	PT83271	80
16	PT86928	80
17	PT72902	80
18	PT92271	80
19	PT94953	80

11). Rank patients by blood glucose level within each gender group.

```sql
--Rank patients by blood glucose level within each gender group
with blood_glucose_level as(
  select Gender, Blood_glucose_level, Patient_id
  from `psyliq-intern.Diabetes_prediction.Diabetes  2`
group by Blood_glucose_level, Gender, Patient_id
)
SELECT
  Patient_id, Blood_glucose_level, Gender, ROW_NUMBER() Over(Order by Blood_glucose_level DESC) as Rank
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
```

Query results

JOB INFORMATION | RESULTS | CHART | PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	Patient_id	Blood_glucose_level	Gender	Rank
1	PT51740	300	Female	1
2	PT49657	300	Male	2
3	PT63081	300	Female	3
4	PT76623	300	Female	4
5	PT5536	300	Female	5
6	PT57767	300	Female	6
7	PT45322	300	Female	7
8	PT78543	300	Female	8
9	PT36031	300	Female	9
10	PT59205	300	Male	10
11	PT54591	300	Female	11
12	PT65140	300	Female	12
13	PT89757	300	Male	13
14	PT99809	300	Male	14

12). Update the smoking history of patients who are older than 50 to "Ex-smoker."

```sql
--Update the smoking history of patients who are older than 50 to "Ex-smoker."
update`psyliq-intern.Diabetes_prediction.Diabetes  2`
Set smoking_history = "Ex-smoker"
where age > 5.0

```

13). Insert a new patient into the database with sample data.

```
1  --Insert a new patient into the database with sample data.
2  insert into `psyliq-intern.Diabetes_prediction.Diabetes  2`
3  values("Jordin", PT5654, Female,5.0,0,0,"Ex-smoker", 22.65,9,155,0)
```

14). Delete all patients with heart disease from the database.

```sql
--Delete all patients with heart disease from the database
DELETE from`psyliq-intern.Diabetes_prediction.Diabetes  2`
where Heart_disease <> 0

```

15). Find patients who have hypertension but not diabetes using the EXCEPT operator.

```sql
-- Find patients who have hypertension but not diabetes using the EXCEPT operator.
SELECT patient_id, hypertension, diabetes
FROM `psyliq-intern.Diabetes_prediction.Diabetes  2`
WHERE hypertension = 1 AND diabetes = 0
```

Query results

JOB INFORMATION | RESULTS | CHART PREVIEW | JSON | EXECUTION DETAILS | EXECUTION GRAPH

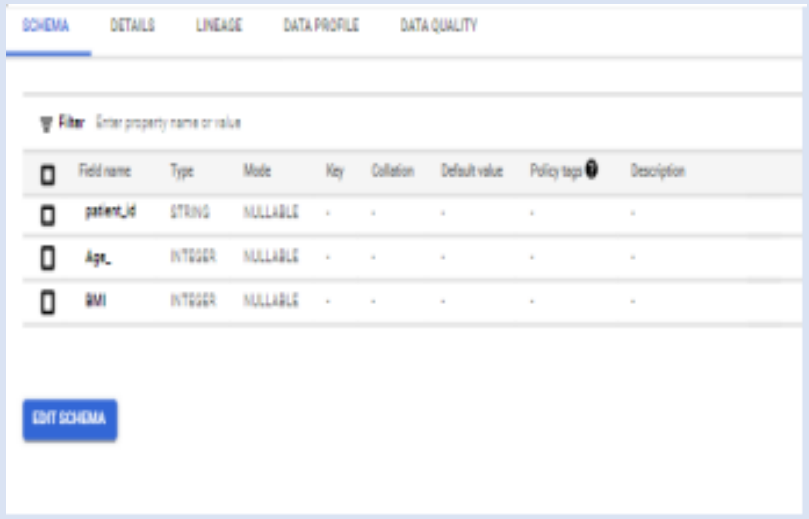
Row	patient_id	hypertension	diabetes
2	PT87841	1	0
3	PT92208	1	0
4	PT59611	1	0
5	PT994	1	0
6	PT14638	1	0
7	PT5413	1	0
8	PT4435	1	0
9	PT74700	1	0
10	PT44273	1	0
11	PT79212	1	0
12	PT24713	1	0
13	PT18824	1	0
14	PT33498	1	0
15	PT80637	1	0
16	PT76198	1	0
17	PT72205	1	0
18	PT1186	1	0
19	PT18676	1	0

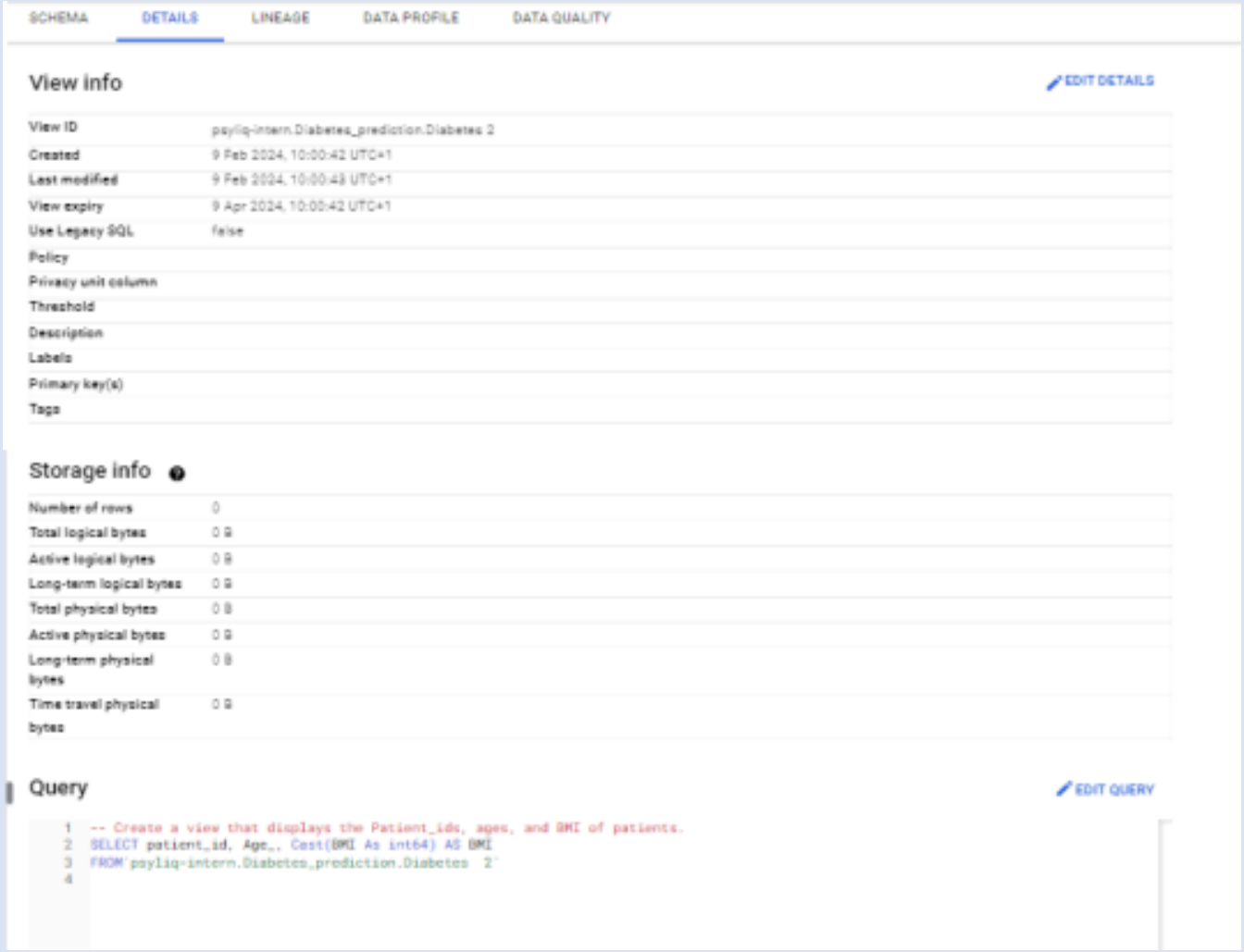
16). Define a unique constraint on the "patient_id" column to ensure its values are unique.

➔ A unique constraint on the Patient_id column would be to ensure its not Null to be sure its values are unique.

1.17). Create a view that displays the Patient_ids, ages, and BMI of patients.

➔ After running a query, click the **Save view** button above the query results window to save the query as a view.
In the **Save view** dialog:
For **Project name**, select a project to store the view.
For **Dataset name**, choose a dataset to store the view. The dataset that contains your view and the dataset that contains the tables referenced by the view must be in the same location.
For **Table name**, enter the name of the view.
Click **Save**.





18).Suggest improvements in the database schema to reduce data redundancy and improve data integrity.

➔ The suggestions below can help improve the database schema to reduce data redundancy and improve data integrity:

- Data integrity mechanisms are essential for reducing data redundancy. By implementing constraints and triggers, you can prevent duplicate data from being entered, enforce referential integrity, and streamline data retrieval thereby ensuring data accuracy, consistency, and validity.
- Also, regular reviewing and updating data integrity rules as your data evolves, helps to maintain data quality and prevent inconsistencies.

19). Explain how you can optimize the performance of SQL queries on this dataset.

To optimize the performance of SQL queries on this dataset, the following can be done:

- Minimize the use of subqueries.
- Avoid unnecessary data retrieval.
- Retrieve only necessary columns.
- SELECT fields instead of using SELECT *.

Thank You