Cloud SQL Instance Setup and Security Documentation

Introduction:

This documentation guides users through the process of setting up a Cloud SQL instance on Google Cloud Platform. It covers essential steps, including instance creation, security configurations, and data analysis.

SetUp Google Cloud Sql Instance:

- Navigate to the Google Cloud Console.
- On the left side of the navigation pane, select "SQL".
- Click the "Create instance" button.
- Choose the database engine (PostgreSQL).
- Configure instance details.
- Instance ID : asmi_psql
- Password : < >
- Database version: PostgreSQL 13
- Cloud Edition : Enterprise
- Choose a preset for this edition: development
- Choose region and zonal availability:
- Region: us-central1
- Zonal availability: single zone
- Customize your instance:
- Machine configurations:
- Machine shapes: 2vCPU, 8 GB
- Storage : storage capacity: 10 GB
- Click "Create instance" to create the instance.

Configure Instance Security

Go to the "Connections" Tab:

- Click on the "Connections" tab to access settings related to network connections.
- Allow only SSL connections
- Only allows connections using SSL/TLS encryption. Certificates will not be verified.

Cloud SQL Connector for Security

IAM Authorization

- The Cloud SQL connector libraries use IAM permissions for access control, allowing you to define who and what can connect to your instances.

Improved Security

- Benefit from robust TLS 1.3 encryption and identity verification between the client connector and the server-side proxy, ensuring secure connections independent of the database protocol.

Convenience

- The connectors eliminate the need to manage SSL certificates, configure firewalls, or handle specific source/destination IP addresses, simplifying the connection process.

Database Schema:

The relational schema designed for this project is based on the provided CSV dataset, named adult.data. The schema includes the following tables:

- **Adult:**
    - age

- workclass
- fnlwgt
- education
- education_num
- marital_status
- occupation
- relationship
- race
- sex
- capital_gain
- capital_loss
- hours_per_week
- native_country
- income

Data Ingestion Process

Python Code Overview

The data ingestion process involves using Python along with the Cloud SQL Python Connector library to import the adult.data dataset into the Cloud SQL instance.

Cloud SQL Python Connector

The Cloud SQL Python Connector library is utilized for connecting and interacting with the Cloud SQL instance securely.

Data Ingestion Code:

```
PROJECT_ID = "york-bb-cohort"

REGION = "us-central1"

INSTANCE_NAME = "asmi-psql"
```

```
INSTANCE_CONNECTION_NAME = f"{PROJECT_ID}:{REGION}:{INSTANCE_NAME}"

INSTANCE_CONNECTION_NAME
```

```
DB_USER = "postgres"

DB_PASS = "password"

DB_NAME = "income_adult_data"
```

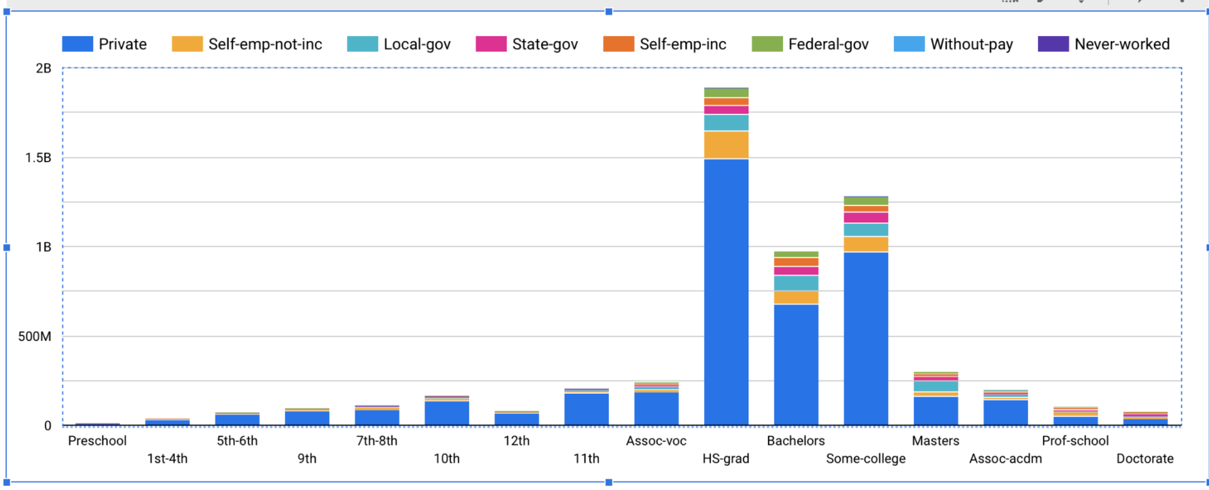
```
from google.cloud.sql.connector import Connector
import sqlalchemy
# initialize Connector object
connector = Connector()

# function to return the database connection object
def getconn():
    conn = connector.connect(
        INSTANCE_CONNECTION_NAME,
        "pg8000",
        user=DB_USER,
        password=DB_PASS,
        db=DB_NAME,
    )
    return conn

# create connection pool with 'creator' argument to our connection object function
pool = sqlalchemy.create_engine(
    "postgresql+pg8000://",
    creator=getconn,
)
data.to_sql('adult', pool, index=False, if_exists='replace')
```

SQL scripts used for analysis:

The relationship between education levels and work classes.

```
SELECT education, workclass, education_num,
sum(fnlwgt) as count_people
FROM `york-bb-cohort.asmi1_dataset.adult`
where workclass is not null
GROUP BY education, workclass,education_num
ORDER BY count_people desc;
```

Education and Work Class



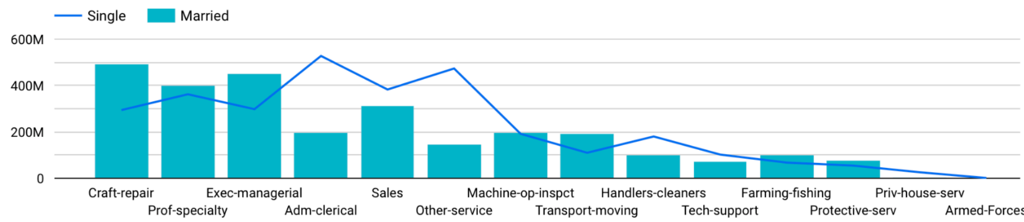
From the above visualization, I infer that most of the working class has at least completed HS-grad.
Also it is evident that most of the jobs are offered by private sector.
It also clearly demonstrates the fact that the majority of working class does attend some form of college and not advanced degree
Another interesting fact is that, the Government Jobs (Federal, State and Local) dwarfs Private Sector jobs by a huge margin.

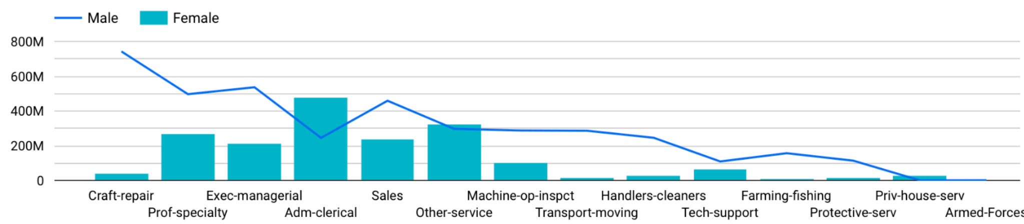
https://lookerstudio.google.com/reporting/f23ddc1b-c0cd-495b-885e-567bea744fe4

Variations in marital status across different occupations

```
SELECT occupation,marital_status,sex,
sum(fnlwgt) as count_people
FROM `york-bb-cohort.asmi1_dataset.adult`
WHERE occupation IS NOT NULL
GROUP BY occupation, marital_status,sex
ORDER BY count_people DESC;
```

Occupation and Marital Status



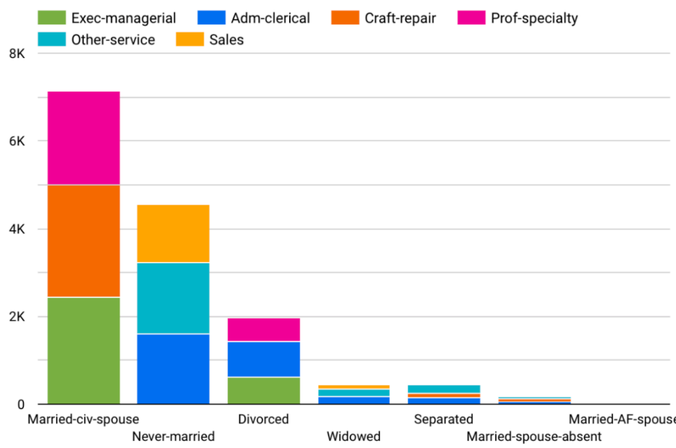


I have grouped individuals who are never-married, separated, divorced, or widowed as single, and categorized all other individuals as married. I found that single individuals are more likely to hold administrative-clerical or other services jobs compared to married individuals. Conversely, married individuals tend to prefer craft-repair and excellent-managerial roles. Additionally, based on separate analysis, it appears that females are more likely to occupy administrative-clerical positions. **Therefore, it can be inferred that single females mostly prefer administrative-clerical jobs.**

https://lookerstudio.google.com/reporting/1514d7a4-b413-4579-9a33-29c116178a34

```
WITH RankedOccupations AS (SELECT marital_status,occupation,
COUNT(*) AS occupation_count,
ROW_NUMBER() OVER (PARTITION BY marital_status ORDER BY COUNT(*) DESC) AS
OccupationRank
FROM `york-bb-cohort.asmi1_dataset.adult`
WHERE occupation IS NOT NULL
GROUP BY marital_status, occupation)
SELECT marital_status,
 occupation,
 occupation_count
FROM
 RankedOccupations
WHERE
 OccupationRank <= 3;
```

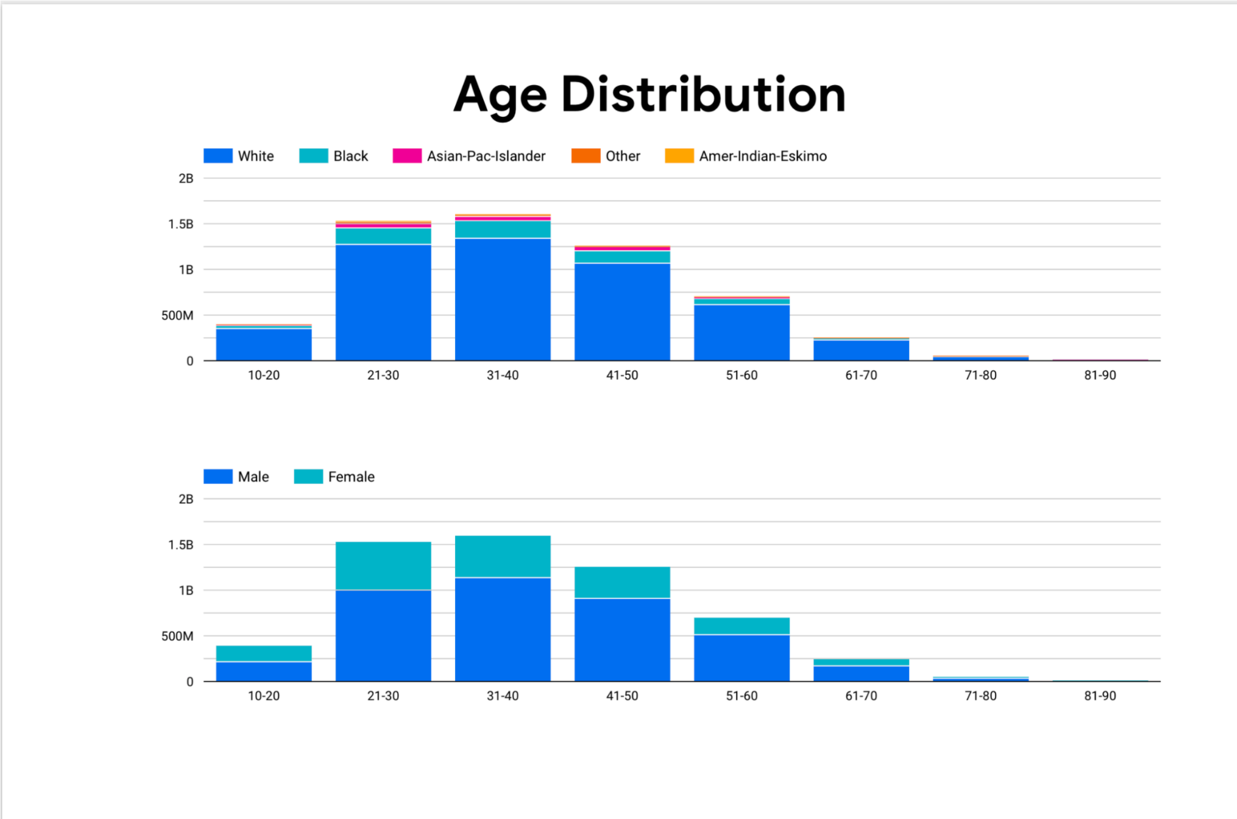
Top occupation count for each marital status.



https://lookerstudio.google.com/reporting/fe320324-e9d1-45b1-a09c-2afd2c2a4182

Age distribution patterns within various racial and gender groups.

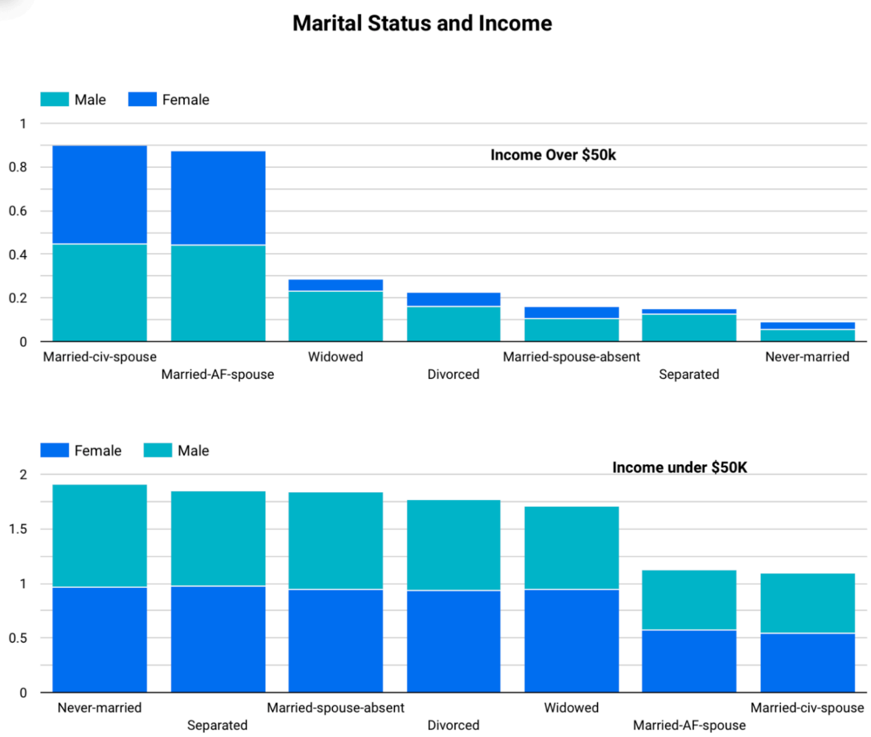
```
age,sex,race,income,
sum(fnlwgt) as count_people
FROM `york-bb-cohort.asmi1_dataset.adult`
WHERE occupation IS NOT NULL
GROUP BY age,sex,race,income
ORDER BY count_people DESC;
```



https://lookerstudio.google.com/reporting/7e2068ab-ef0e-48e6-b9a0-7d66b4af8989

Influences of certain demographic features on marital trends.

```
SELECT marital_status,sex,
   AVG(CASE WHEN income = '>50K' THEN 1 ELSE 0 END) AS avg_income_over_50k,
   AVG(CASE WHEN income = '<=50K' THEN 1 ELSE 0 END) AS avg_income_under_50k,
   sum(fnlwgt)as count_people
FROM `york-bb-cohort.asmi1_dataset.adult`
GROUP BY marital_status,sex
order bycount_people;
```

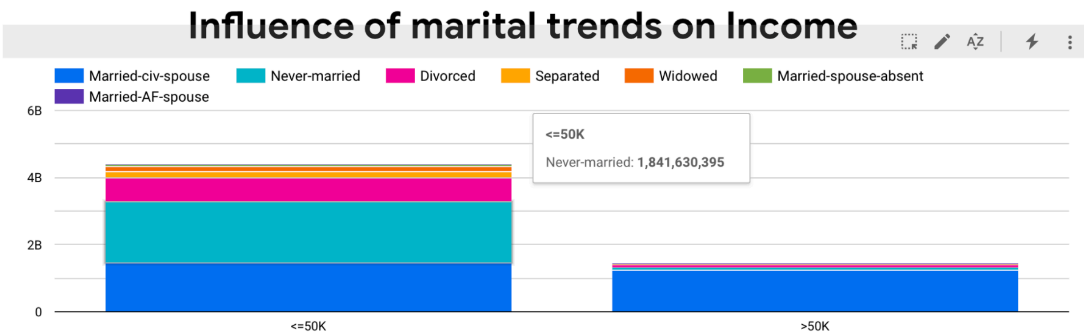


Income over $50K is the same for married males and females, but single males have a higher proportion. Income under $50K is equal between married and single males and females.

https://lookerstudio.google.com/reporting/95abf372-66c1-447d-baad-43a51e89e568

```
SELECT sex,marital_status,education,income,sum(fnlwgt) as count_people,
```

```sql
FROM `york-bb-cohort.asmi1_dataset.adult`

WHERE occupation IS NOT NULL

GROUP BY sex,marital_status,education,income

ORDER BY count_people DESC;
```



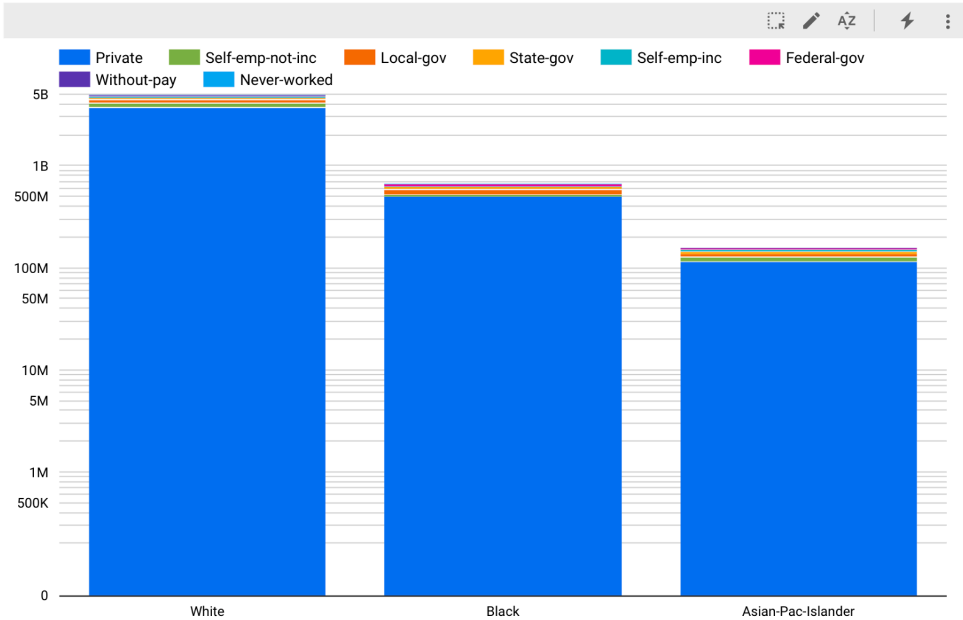
This chart shows most of the Never married people make under 50k and the reason could be they are early-on in their career. But for married people it doesn't make a big difference as they can choose between under or above 50k jobs.

https://lookerstudio.google.com/reporting/b6fda097-7249-49ad-bf21-d3de51fca98e

Employment trends analysis by race

```sql
SELECT race,workclass,
sum(fnlwgt) as frequency
FROM `york-bb-cohort.asmi1_dataset.adult`
where workclass is not null
GROUP BY race, workclass
ORDER BY race, frequency DESC;
```

Race and Work class



Note: For better Viewing, I have filtered Race Types Other and American- Indian-Eskimo. Also I have changed the y axis to Log Scale.

From the above chart it is clearly evident that White is Majority among all work classes followed by Black.

https://lookerstudio.google.com/reporting/f3dc2527-6624-4555-8ec5-3c2d03e8b92f