Data Warehouse

HHUSA Dataset

Sanchayan Bhunia

About The Problem

- HHUSA Dataset from Teradata

- Answer questions related to Employer Partnerships and Opportunities and Serving Spouses Program.

- Employer Partnership and Opportunities
  - Effectiveness of email programme in terms of job seekers' profile creation
  - Temporal analysis of number of accounts (both job seekers and employers)

- Serving Spouses Program
  - Spouse spent days in HHUSA programme.
  - Veteran spent days in HHUSA

- Spouse Demographic Profile (Gender, Location, Education etc.)
- Service Member Status vs Spent time in HHUSA
- Co-relation in Education and Date status turned black
- Spouse Employed vs Unemployed
- How many Blue (hired) vs Grey (inactive)

Observation

- Contact contains contact information
  - If a client or partner
  - If a veteran or spouse
  - Education
- Account contains information about account creation date
- Contact might or might not have an account
- Contact Creation **Date** can be way in advance
- Email history file has contact id
- In some cases Date hired > Contact creation date

Assumptions

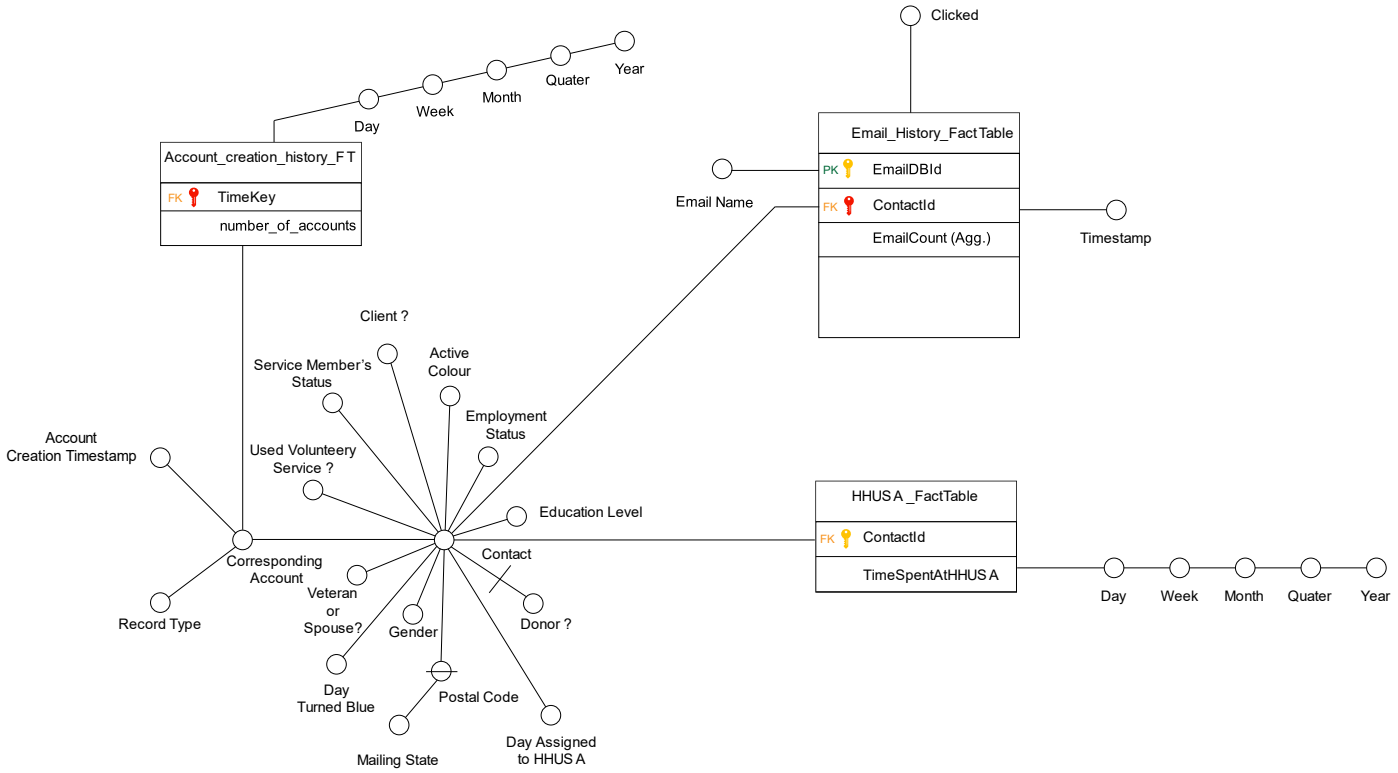
- Job seekers are clients
- Employers are partners
- Contact table has both of them, just clients are flagged 1
- Spouse are flagged 1
- “Top jobs” and “virtual career fare” marked emails are important
- Time spent at HHUSA = date assigned to hhusa – time of hire(blue)

Cleaning and Integration

- Cleaning and integration are on Tableau Prep Builder
- Select contactID, AccountId, Gender, Education, member status etc. (Contact)
- Select Contact_id, AccountID, Creation Date etc.(Account)
- Email_DB_name, name, sent_date and filter “job fair” etc.
- Create CSV for spouse, veteran, partner, email, account
- Ready to load in server

DFM

- Three Fact Tables
- Three measures in them
- Contact Dimension shared by two FTs
- Account dimension shared by one FT
- Donor is optional
- Mailing state incomplete hierarchy
- Issue solved in ETL



Logical Design Choices

- Data Warehouse is focused toward answering the questions.
- Requirements:
  - Email efficiency -- New client accounts created after sending email
  - Accounts – Analyse account creation pattern in hour, month, year, quarter
  - Contacts – should keep link to account and email
  - Low on Workload
- A lot of analysis on spouse
- Less analysis on veteran
- Almost none partners so, we will split partners and keep only their contact_id

Workload Approximation

- Two choices
  - Keep client contact intact
  - Split the contact in spouse and veteran

- Lets calculate the workload if we split the data

$$The\ volume\ of\ all\ the\ events\ related\ to\ our\ queries = 15 \times 106000 = 1590000$$

$$Let\ say\ that\ that\ is\ the\ total\ frequency = 1$$

Workload without splitting = 1590000

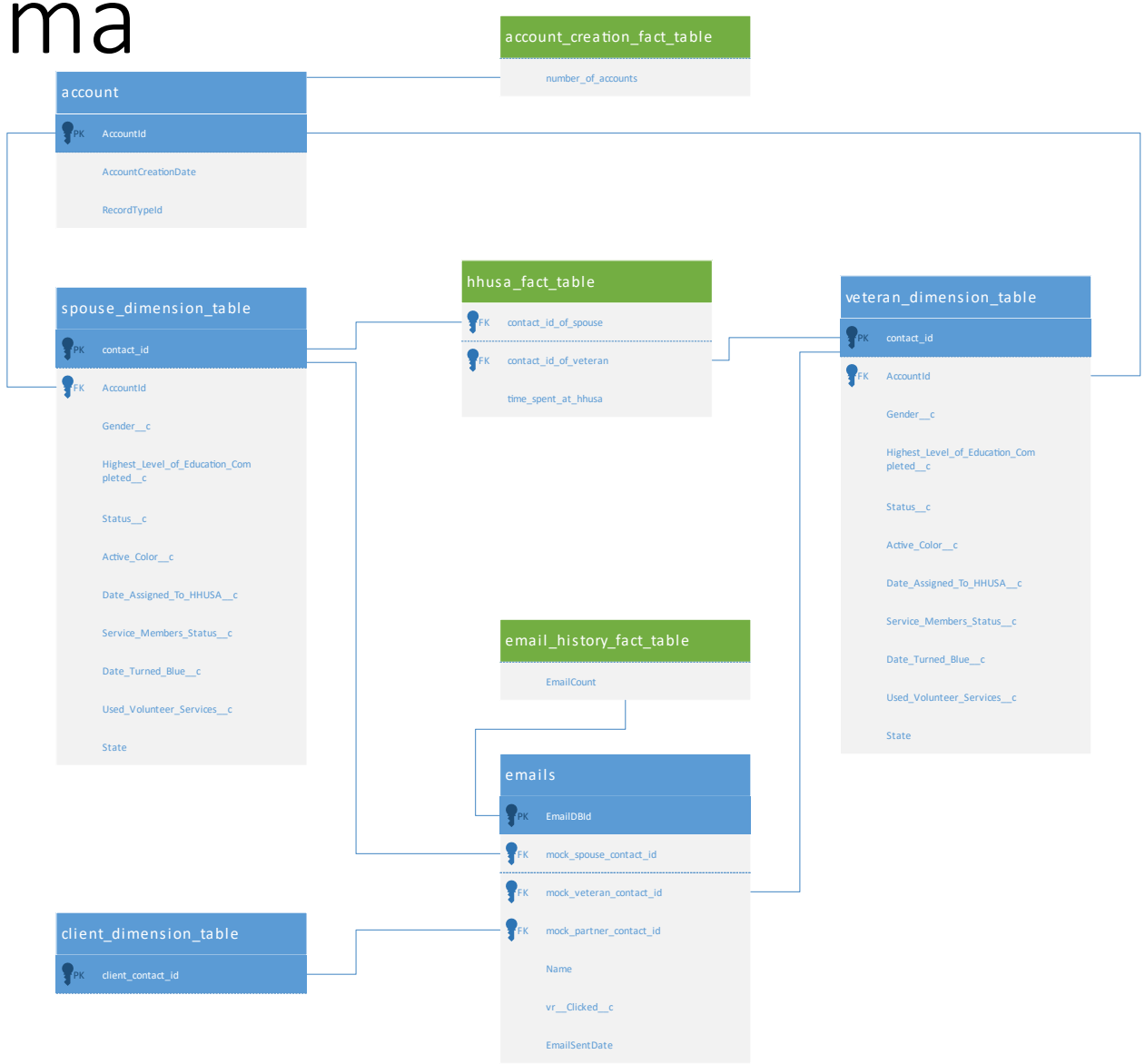
$$Volume\ of\ events\ related\ to\ only\ spouse\ occupy\ a\ volume = 75000$$

$$Frequency = \frac{6}{9} = 0.67$$

$$So, now\ Workload = \{(1 - 0.67) \times (1590000\ -\ 75000)\} + (0.67\ \times 75000)$$

$$= 550200$$

ROLAP Schema



ETL

- Tableau Prep is used for extraction and transformation of the data
- In order to deal with the inconsistent **location** information we join an external file containing full description of US states.
- Save the tables in CSV format
- Loading:
  - Create tables with corresponding column names
  - Set primary and foreign keys
  - Load the csv data into dimension tables

```
COPY table_name FROM '/path_to_csv_file.csv' DELIMITERS ',' CSV;
```

  - Populate the Fact tables

Sample OLAP Queries

- **Yearly Performance of Quarters**

```
SELECT
date_part('quarter', account.account_creation_date) as quarters,
date_part('year', account.account_creation_date) as years,
COUNT(account.account_id)
FROM account
GROUP BY quarters, years
ORDER BY quarters, years;
```

- **Rolling average of account creation in hours of a day over all years**

```
SELECT hours_of_day.*, AVG(counts)
OVER (PARTITION BY day ORDER BY hour rows between 11 preceding and current row) as
rolling_average FROM
(SELECT  date_part('hour', account.account_creation_date) as
        hour, date_part('day',
        account.account_creation_date) as day,
        COUNT(account.account_id) as counts
FROM account
GROUP BY day, hour
ORDER BY hour) hours_of_day
```

- **Mobile average of time spent by spouse on their highest degree**

```
select
 contact.highest_level_of_education_completed__c, contact.time_spent_at_hhusa,
avg(contact.time_spent_at_hhusa) over(order by contact.highest_level_of_education_completed__c ROWS
2 PRECEDING)
from
(
select contact_id_of_spouse, time_spent_at_hhusa, spouse_dimension_table.state,
spouse_dimension_table.highest_level_of_education_completed__c
from hhusa_fact_table
inner join spouse_dimension_table
on spouse_dimension_table.contact_id = hhusa_fact_table.contact_id_of_spouse
where time_spent_at_hhusa >0
) contact
group by contact.highest_level_of_education_completed__c, contact.time_spent_at_hhusa
```

Outcomes

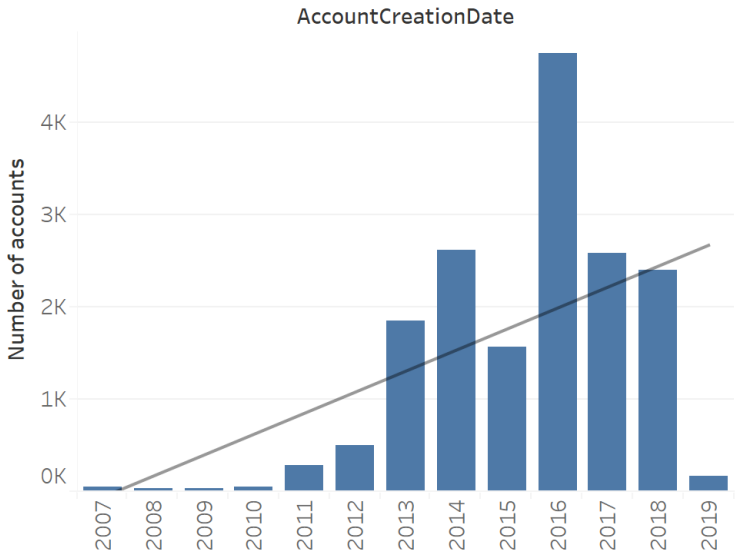
- **Effectiveness of Email Campaign**

$$\frac{\textbf{\# of new Client Accounts}}{\textbf{Total \#of job emails sent} - (\textbf{\#of emails sent to partners} + \textbf{\#of accounts already existed before email was sent})}$$

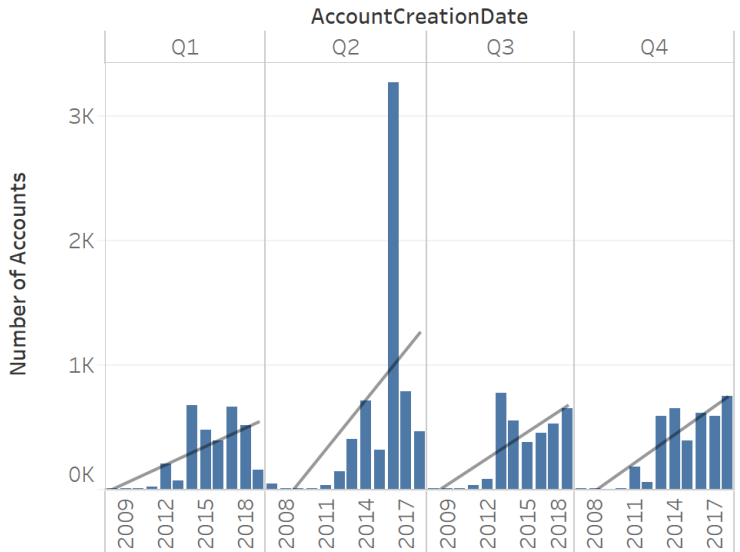
$\approx 3\% \ (Not\ Bad!)$

Yearly performance of quarters an uptrend in all

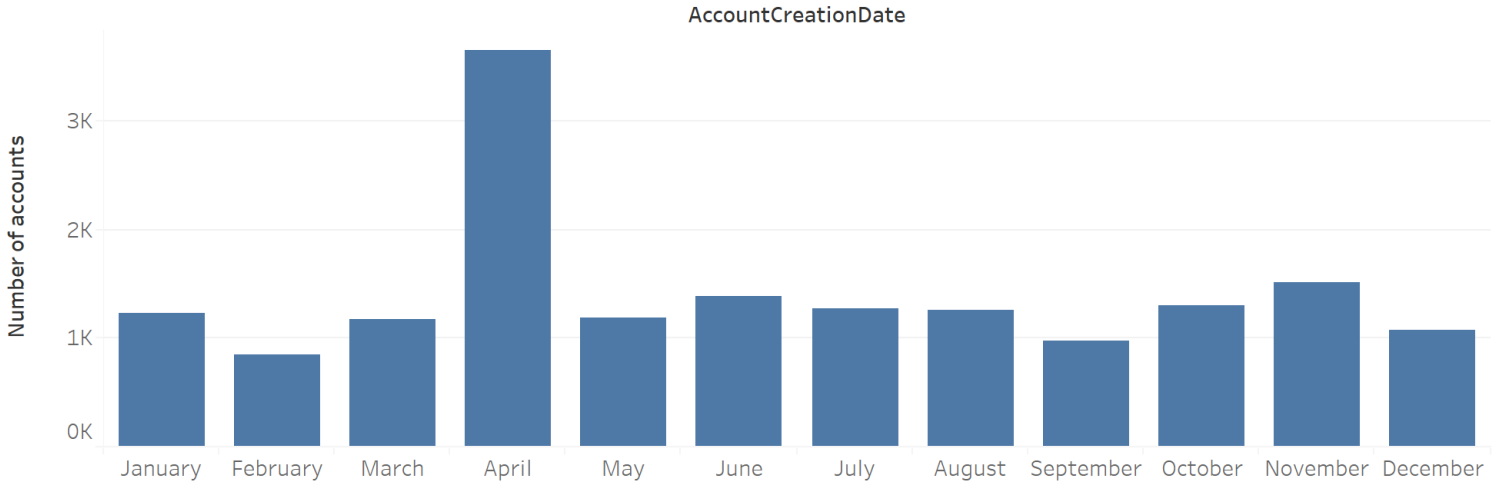
Yearly Analysis



Yearly Performance of Quarters

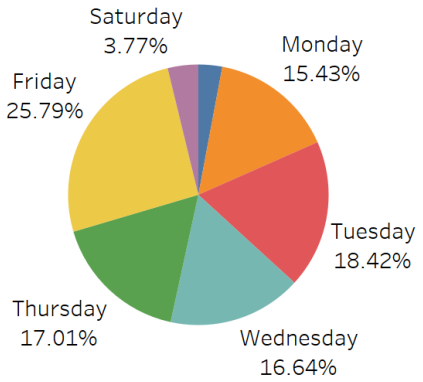


Monthly Analysis



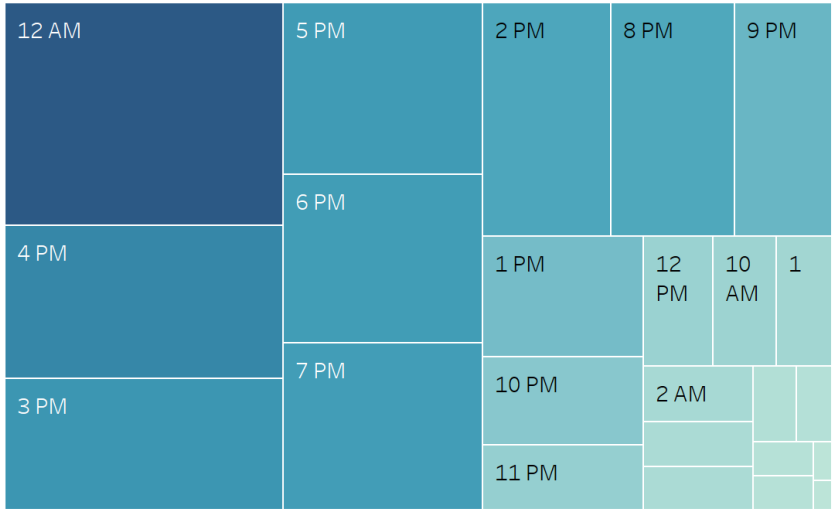
Hourly and Weekly analysis

Performence of days in week



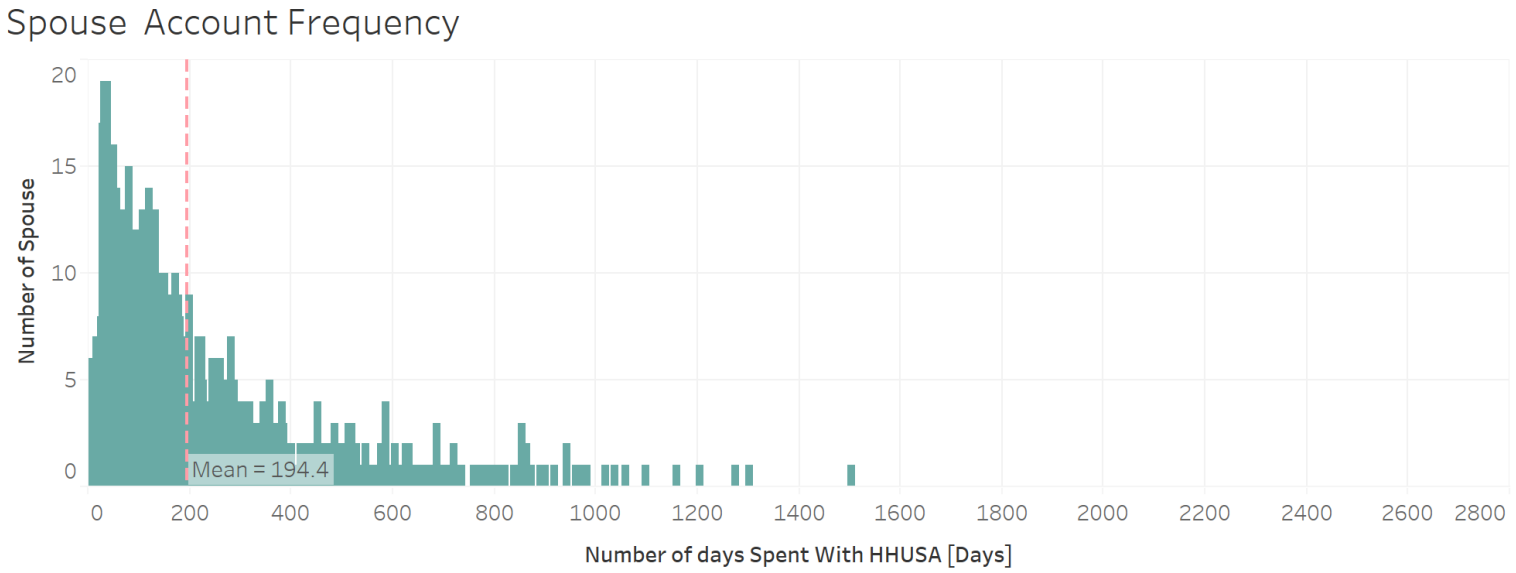
Weekday of AccountCreati..
- Sunday
- Monday
- Tuesday
- Wednesday
- Thursday
- Friday
- Saturday

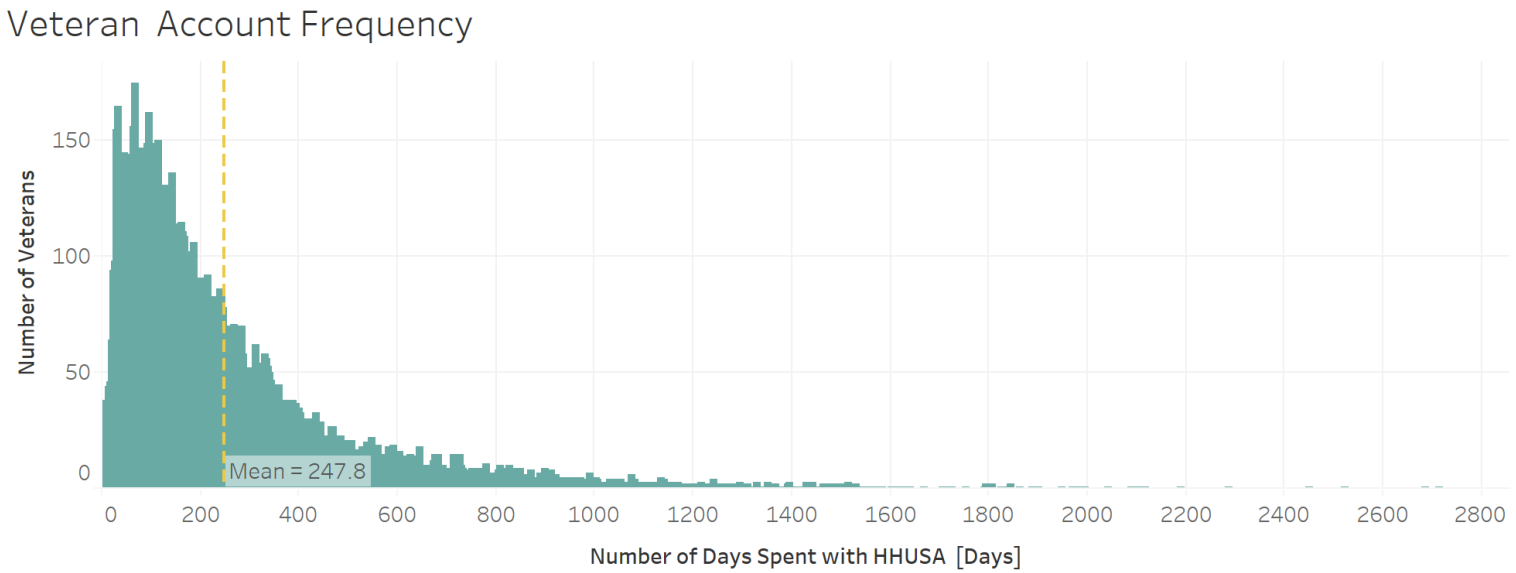
Hours in a day



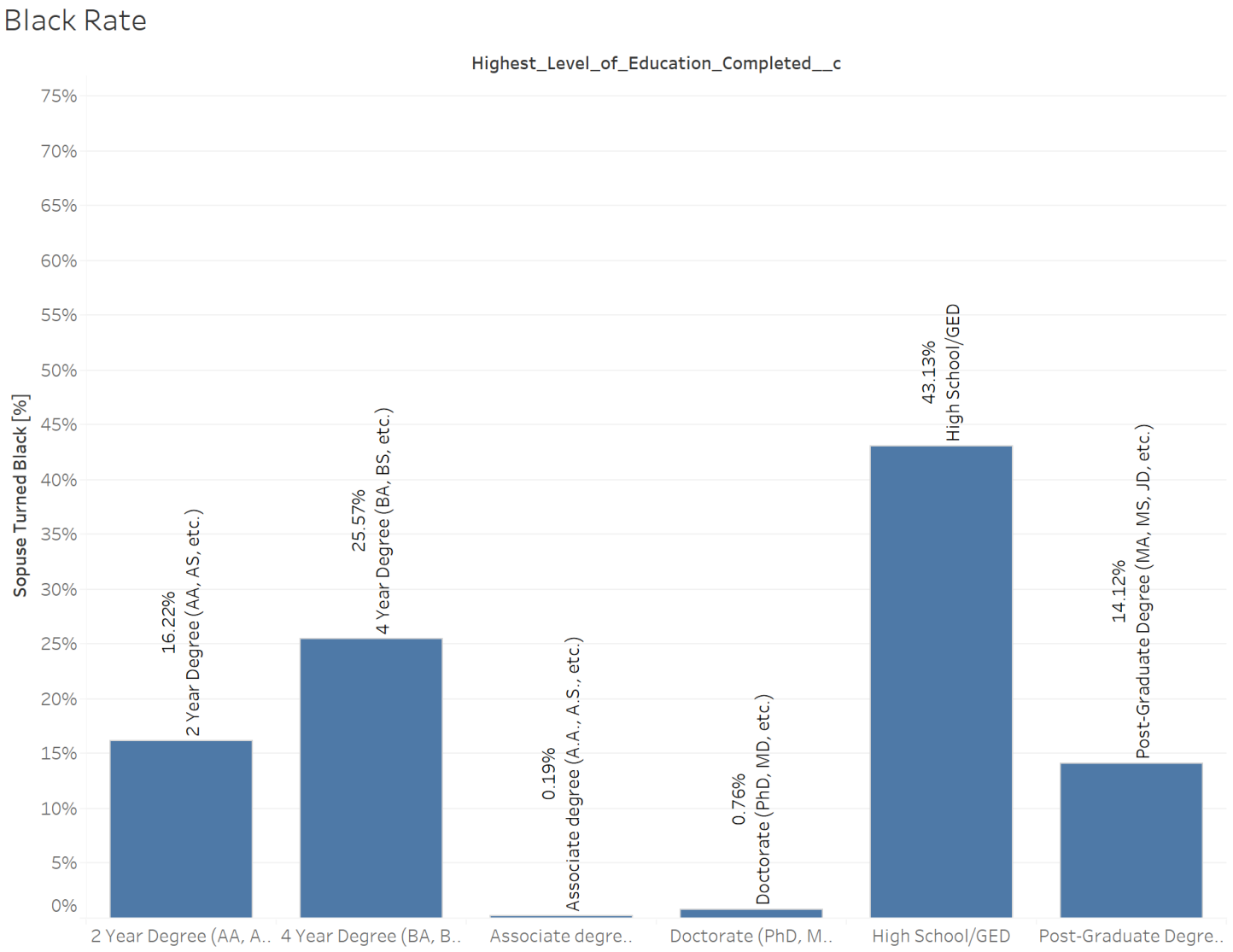
Count of AccountId
22 2,500

Number of days spent on HHUSA

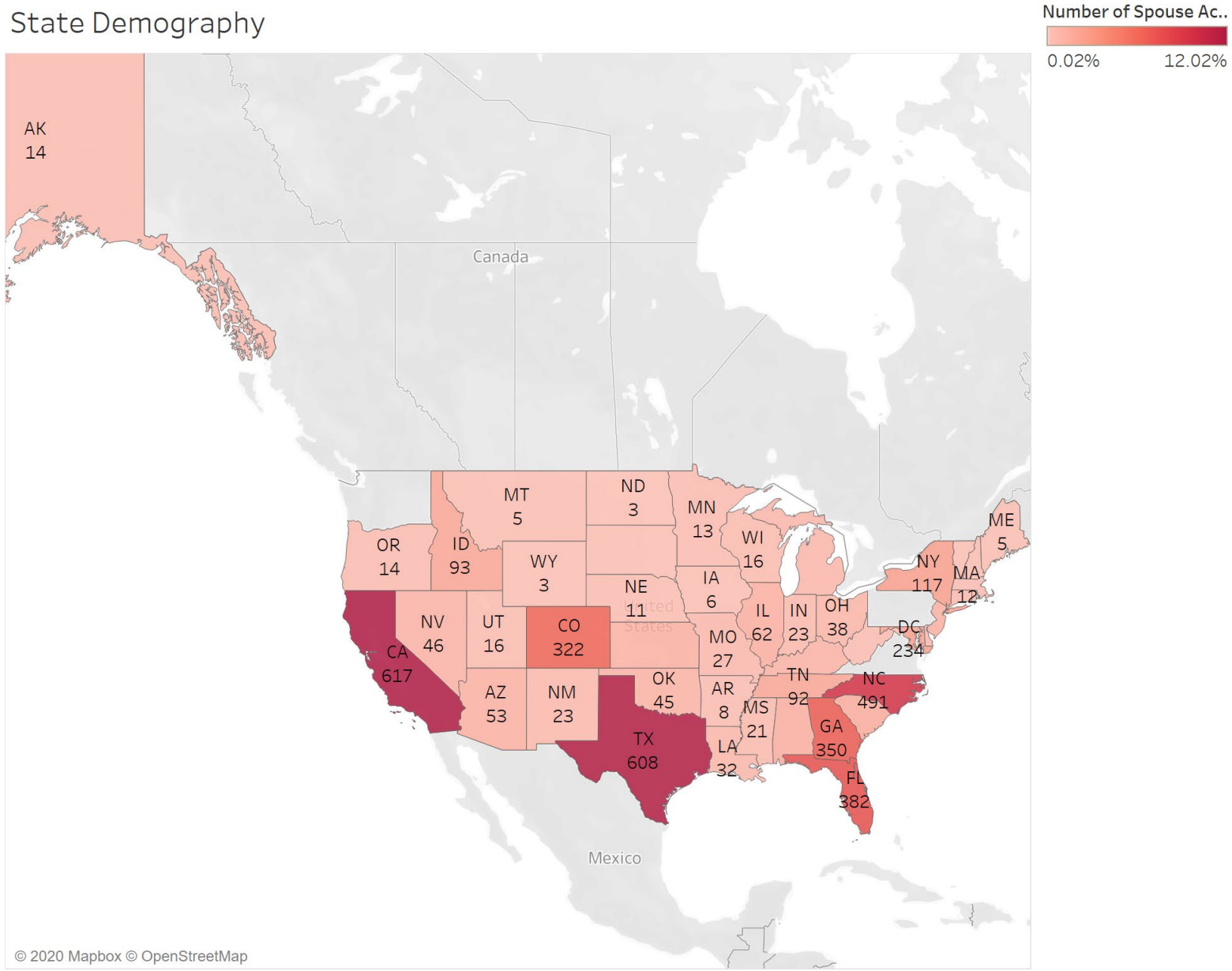




Relationship between spouse turning black and education level



Spouse account frequency from different states



Estimated Effort

70 hrs(approx)

- Due to none to very little documentation about the data source, the inspection and profiling took almost 20hrs.
- Cleaning and integration took around 5 hours.
- DFM conceptual modelling 5hrs.
- Logical Modelling took around 10hrs (tried different models with workloads).
- ETL Process took around 6hrs.
- Queries took around 5hrs.
- Analysis took 10hrs.
- Documentation around 4hrs.