6/4/2020

HHUSA Project Documentation

Data Warehouse Project

Sanchayan Bhunia
UNIVERSITY OF GENOVA

HIRE HERO USA DATA WAREHOUSE DOCUMENTATION

May 30, 2020

OVERVIEW

1. Introduction

Hire Hero USA is a non-profit organization devoted towards connecting veterans and their spouses with organizations willing to hire. Not only that, they also provide volunteer services to willing participants by conducting mock interview sessions and other necessary arrangements while logging their data in transactional data bases. Such a data source is been analysed here in order to find relevant correlations in Employer Partnerships and Opportunities and Serving Spouses Program guided by the business questions in Teradata.

2. Scope of The Data-Warehouse Design

The scope of the data warehouse is to answer some of the important business questions related to Employer Partnerships and Opportunities Serving Spouses Program and form Hire Hero USA data. The specific questions that need to be answered from both of the fields are the following.

Employer Partnerships and Opportunities

Q: Do email campaigns have any effect on job seekers creating profiles on the Hire Heroes USA Job Board?

Q: Is there a relationship between certain days of the week, times in the day, or months, or time of year and when employers and job seekers create accounts?

Serving Spouses Program

Q: What is the average amount in days that a military spouse spends in the HHUSA program?

- How does this differ from the average veteran client?
- What is the demographic profile of the military spouses who are registering (gender, location, service members status, level of education?)
- How does their Service Members status affect their time to hired status?

Q: Is there a correlation between education level and the black rate for military spouses.

- What percentage of military spouse clients say that they are underemployed vs. unemployed?
- How many spouses who registered for services who say that they are underemployed turn blue vs grey?

3. Data Source Observations & Assumptions

Assumptions are made by inspecting the dataset after importing them into Tableau Prep Builder. The notable assumptions after observations are the following.

Observations

The file contacts contains all different attributes regarding contact like contact id, account id, contact creation date, contact is a client or not, gender, if the contact is a veteran or a spouse of a veteran, contact address, if it is a donor, if it is a job seeker, their job status, previous status and so on. On the other hand, the accounts file contains information about the account id, account creation date, if the account is a partner account or a client account and so on. A contact might or might not have an account but not the other way around. And it is observed that a contact creation date might be years way before the account was created.

The email-history file contains the information about the contact to whom the email was sent, the content of the email, the email sent date etc.

After close inspection we also find that in many cases, the dates they got hired (blue status) are surprisingly years ahead of their account creation dates on job desk and sometimes even before the contact creation dates.

Assumptions

- The way it works is that HHUSA already has a made a list of contacts from either from their previous campaigns or from different data sources and government records, then they send them emails or call the contacts then interested contacts creates an account on their job board.
- The same list of contacts also contains the information about their partners.
- Clients are job seekers.
- Partners are employers from different corporations.
- Donors Can be both partners and Clients.
- Account table contains both partner and client accounts.
- Contact table contains both partner and client accounts but only clients are flagged (1).
- Contact creation datetime is always before account creation datetime.
- Emails identified as “Top Jobs” or “Virtual Career Fair” in Targeted Email Subject are related to opportunities
- All contacts those have an account could be from a client or a partner could be found from inner join of Contacts and Accounts.
- The clients can either be a veteran or a spouse of a veteran.
- For calculating a client’s time spent on HHUSA programme, we take the time difference between Dates_Assigned_to_HHUSA and the date when their status turned blue referring to a hired status in order to avoid nonsensical behaviour of the given data set mentioned in the observation stanza.

4. Cleaning and Data Integration

The cleaning and integration are done using Tableau Prep Builder software package. After importing four files of our interest, i.e. contacts, emails, accounts and us state names and abbreviation (added from external sources to get us state names) we can follow the tableau prep flow in order to get our desired outputs.

Cleaning

The contact file consists of many attributes and but we will need only ContactId, AccountID, ContactCreationDate, Client__c for employer partnership part and ContactId, MailingState, Gender__c, Highest_Level_of_Education_Completed, Staus__c, Date_Assigned_to_HHUSA__c, Service_Members_Status__c, Federal_Program_Participant__c, Military_Spouse_Caregiver__c, Used_Volunteer_Services__c, Date_Turned_Black__c for serving spouse programme. We also fix the lower cases and remove null values in mailing state in order to join them with the usa_state table to get the full description of the states from abbreviation.

In the account file we only need AccountId, RecordTypeId and AccountCreationDate.

In the email table we only need EmailDBId, Name, EmailSentDate, vr__Clicked__c, ContactID.

Integration

Employer partnership and opportunities Analysis related

After deciding what attributes, we require we must clean the dataset in order to integrate them in final step. We are interested in finding which contacts has an account associated with it because those are the contact where we can check if the account was created after sending them email or they already existed before sending email. In order to do that, we filter the clients out by Client__c flag in contacts table and inner join them with account table by accountId and we will get the contacts of only the clients with account namely “job seekers”.

The Partners only accounts are found from subtracting client only accounts from inner join of all contacts and all accounts and set as an output.

Emails were sent to email addresses from both existing contacts and new email addresses not in any contacts (referred by null in the ContactId column of Email_History table). AS a matter of fact, we will simply export out cleaned email_history table as an output file in order to load them in PostgreSQL server.

Serving spouse programme

For this part we need to focus again on the contacts table. Here we can filter out the military spouse accounts by Military_Spouse_Caregiver__c flag true. And the veteran accounts by the same flag false. Alongside that, we also keep in mind to carry on with their gender, their home address, service status and employment status, their educational qualification. But the mailing states mentioned here are not in details but an abbreviation so, we need to integrate them with the usa_state table in order to get the detail description of the states where the contacts are from. Then we can export the output file in order to use them in SQL server or for further analysis in Tableau Desktop application.

5. Conceptual Modelling

> *For conceptual modelling we can either use an ER Model or a DF Model. But as our purpose of analysing the dataset requires us to build Data Marts, Entity-Relationship Model can not be used. Rather we will proceed with our model of choice as the Dimensional Fact Model.*

Dimensional Fact Model

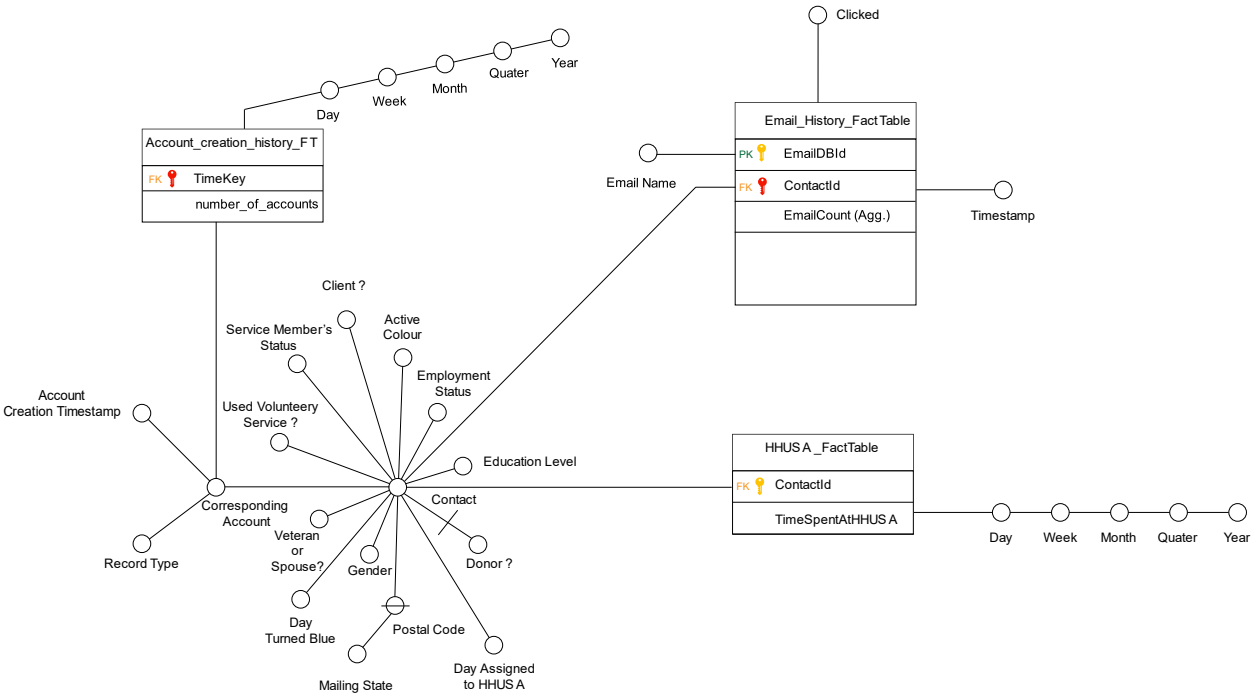
Motivation behind the Design Choice

Based upon the business questions we want to solve, there are three distinct facts to be explored. Emails were sent as a part of campaigns or whatsoever, accounts were created and on the client’s side there is HHUSA programme. So, if we want to design a single Data Warehouse consisting three different Data Marts, we must consider three Fact

Tables. The time recorded in the tables are in timestamp format which can easily be modelled as a hierarchy of primary and secondary dimensions. Alongside that, contacts which has associated accounts can also be modelled as a hierarchy. We have considered if a contact is a donor or not, but that is not relevant for this analysis and marked as an optional dimension in the model. The mailing state dimension is an incomplete hierarchy with respect to postal code and is been marked in the model. Contact dimension is shared between both fact tables making it a confirmed dimension.

There are different measures with respect to each of the three fact tables like total time spent on HHUSA programme, Average time to get hired for clients which must be calculated in HHUSA Fact table, to analyse the trend in number of accounts created we need to aggregate over temporal dimensions and to measure the effectiveness of email campaign we need total number of emails sent to the contacts and check how many of those contacts opened accounts after receiving the email which must be calculate in the Email History Fact Table.

The Model



6. OLAP Logical Design of Data Warehouse

There are too many numbers of ways to design a Data Warehouse for the above conceptual model. Some are expensive in terms of space and required time to execute a query. The point to be kept in mind that, we already know what the questions are we want our warehouse to answer. So, our goal is to minimize the calculated workload by efficiently designing the logical schema from the conceptual model.

Questions regarding Email Campaign efficiency requires to find out the client contacts which did not have an account earlier but created their account after receiving emails regarding job opportunities namely “Top Jobs” or “Virtual Career Fair” in Targeted Email Subject. So, we will filter out only those emails from email database which will help us save both space and query time. Then there are two options either we can filter out the client’s contacts or we can take both clients and partner’s contacts and join it with account database to get the account creation datetime and compare them with email datetime. But in the later part of the question where we answer the temporal queries, we can see that we need both “job seekers” and “employers” i.e. both partners and clients. So, we will keep accounts of both the parties.

For answering questions regarding the Spouse Serving Programme we can go for two strategies, one, splitting the contact database into spouse and veteran tables, two, keeping them intact. And as spouses and veterans are considered as clients, we can filter only clients as we had to do for the previous analysis. Other dimensions do not matter a lot in this case because our queries like time spent at HHUSA and Time to get hired require the measures to keep in days granularity.

Workload Calculation

Now let's calculate the workload for both cases and see if there is any significant benefit over splitting the contacts in terms of spouse and veterans.

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

$$Number\ of\ queries\ that\ will\ access\ the\ volume = 9$$

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

$$So, Workload = 1 \times 1590000 = 1590000$$

$$Volume\ of\ events\ related\ to\ only\ spouse\ occupy\ a\ volume = 15 \times 5000 = 75000$$

$$Number\ of\ queries\ that\ will\ access\ that\ volume = 6$$

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

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

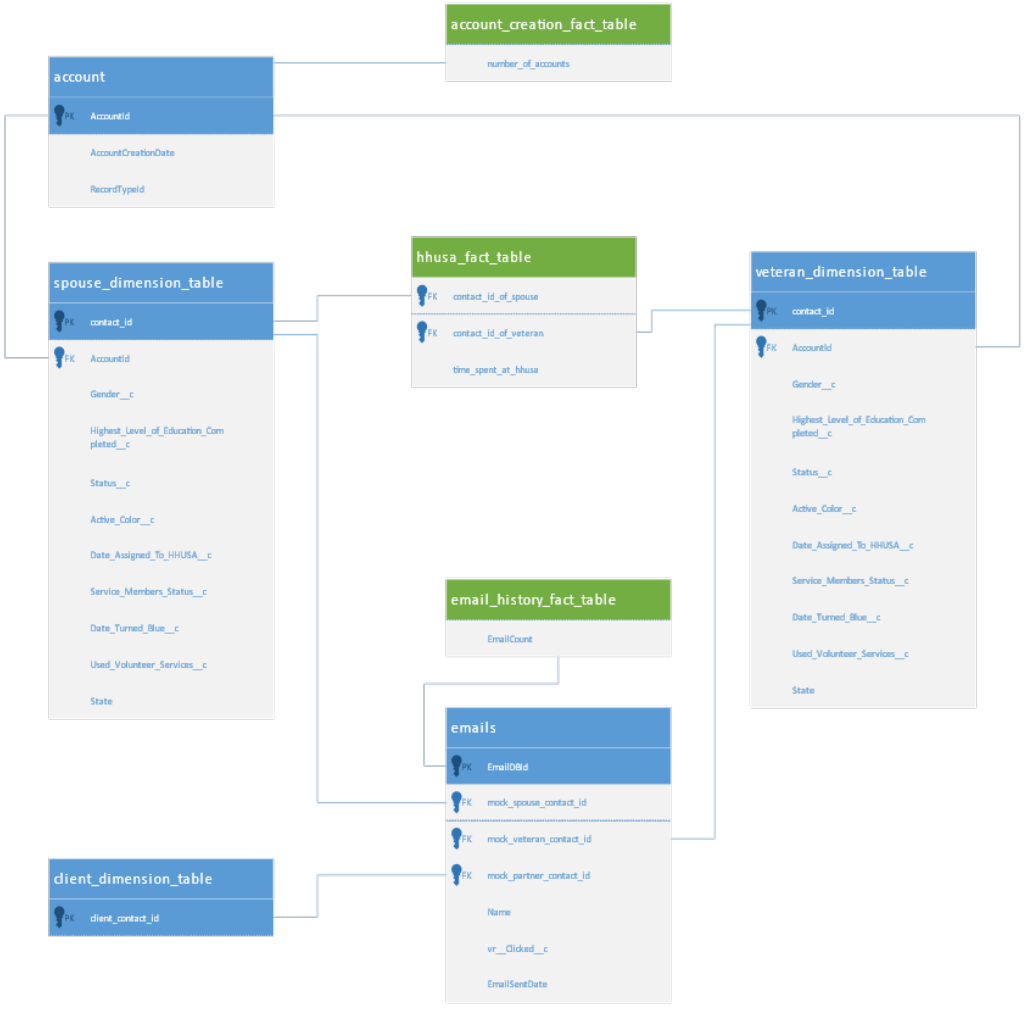
$$= (0.33\ \times 1515000) + 50250$$

$$= 499950 + 50250$$

$$= 550200$$

From the workload calculation, there is a significant benefit in splitting the dataset in terms of veteran and spouse.

7. ROLAP Logical Design



8. ETL Process

Tableau Prep Builder is been used for extraction and transformation of the data from the csv files provided. The loading the data into the database is done by through importing the data corresponding tables in Postgre server through pgAdmin. Most relevant steps are mentioned in the "OLAP Commands.docx" file in Documentation folder.

Important choice here while we create the emails table. Here we have duplicated the contact_id three times for mock_spouse_contact_id, mock_veteran_contact_id and mock_partner_contact_id and made them FOREIGN KEYS in a hope that when we INNER JOIN them with relevant tables, we will get correct contact_id and related email_db_id corresponding to spouse, veteran and partner.

We do not perform any relevant analysis in client's side, so we have just kept the contact_id in client account.

Tableau Prep Builder cannot generate date dimension so we will use date_part () in order to explore the time dimension.

> *For details about the ETL process and to get the exact flow please look into "Employer Partnerships and Spouse Programme.tfl" or "Employer Partnerships and Spouse Programme.tflx" in "Tableau Prep\Flows"*

9. Results

Effectiveness of Email Campaign

> *For the queries related to this calculation please investigate the "OLAP Commands.docx" file*

- Effectiveness of the campaign in terms of job seekers creating accounts can be calculated as what percent of new accounts were created with respect to emails sent to potential clients.
- Not all potential clients to whom the emails were sent, were not on the contact table.
- Job related emails were sent to partners as well but those are irrelevant while calculating the effectiveness.
- Job related emails were also sent to already existing clients and those are irrelevant for effectiveness calculation.
- So, emails sent to potential clients =
  $$\textbf{Total \#of job emails sent} - (\textbf{\#of emails sent to partners} + \textbf{\#of accounts already existed before email was sent})$$

Effectiveness of email programme on job seekers creating profiles =

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

$$= 96/\{19341 - (6626 - 9379)\}$$

$$= 0.02877$$

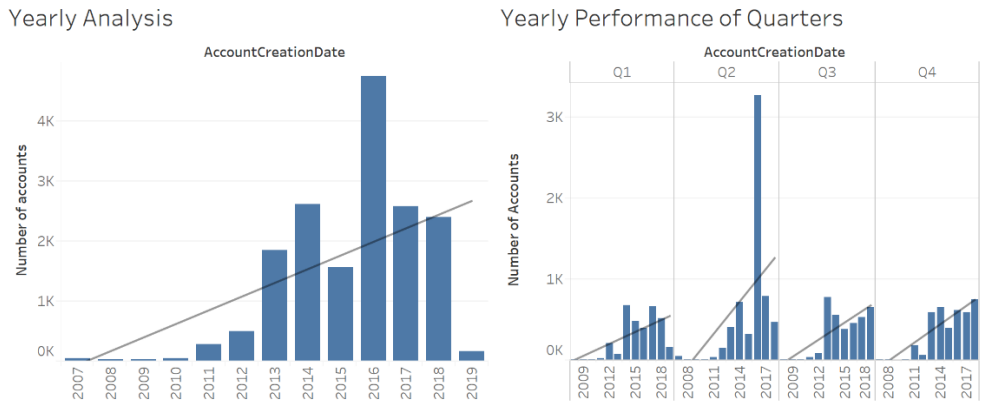
$$\approx 3\% \ (Not\ Bad!)$$

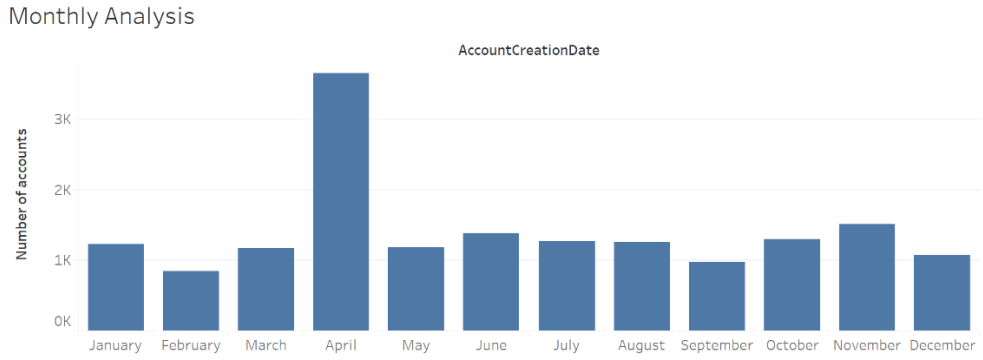
Trends in Account Creation

> *To visualize the results, we will connect Tableau Desktop application with our Postgre Server*

We are interested in finding is there temporal a pattern in clients and partners creating accounts on HHUSA. In order to do that we have to make some queries in the PostgreSQL and we will visualize them using Tableau Desktop application.

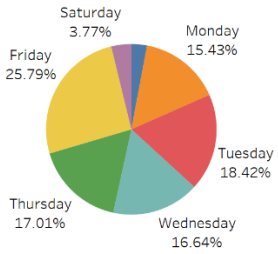
Yearly performance of quarters an uptrend in all

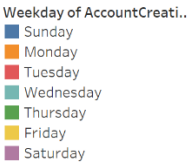




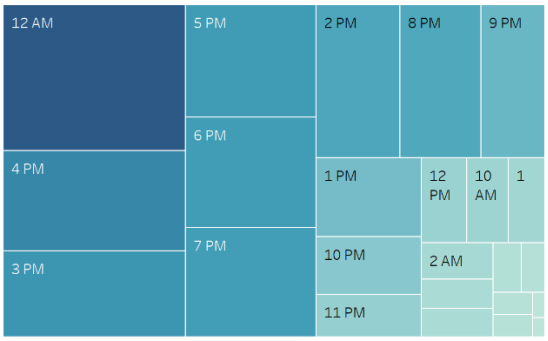
Day in week and hours in day analysis

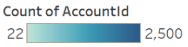
Performence of days in week





Hours in a day

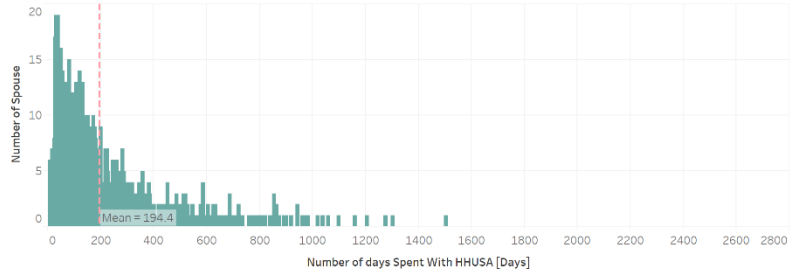




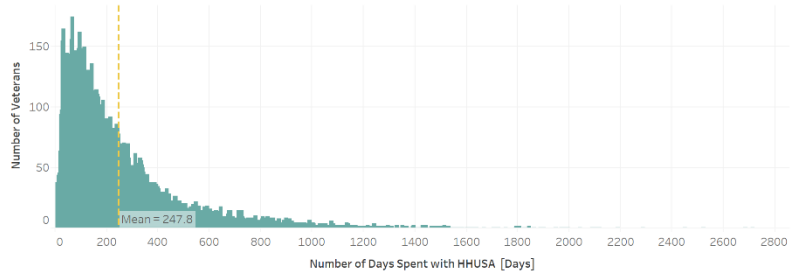
Spouse Programme

Average number of days spent by military spouse compared to veterans

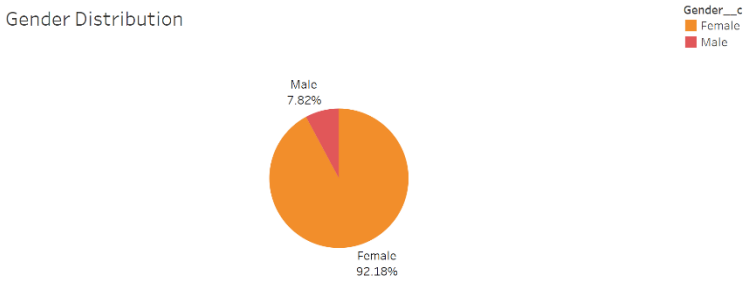
Spouse Account Frequency

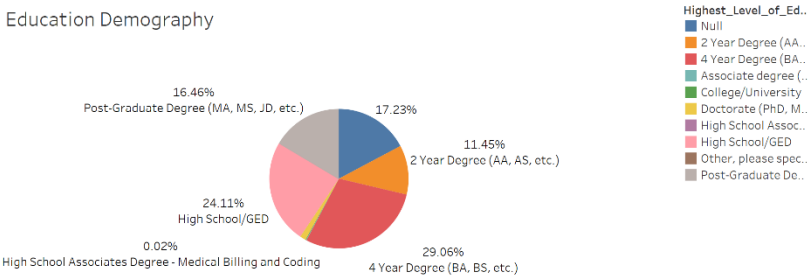


Veteran Account Frequency

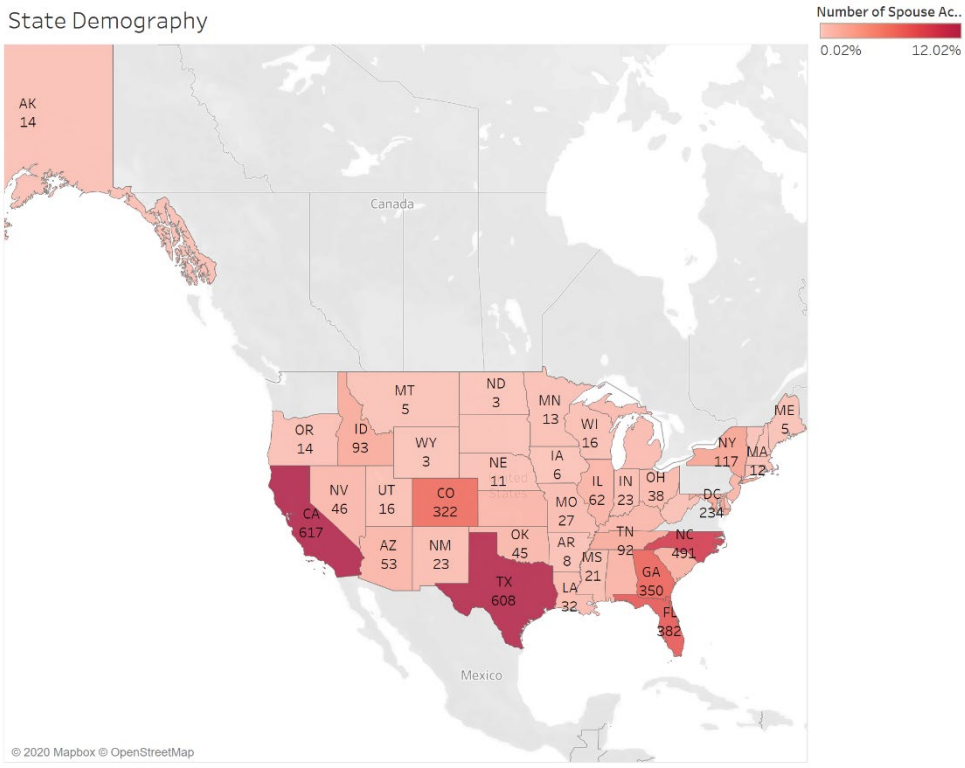


Demographic Profile of Spouse

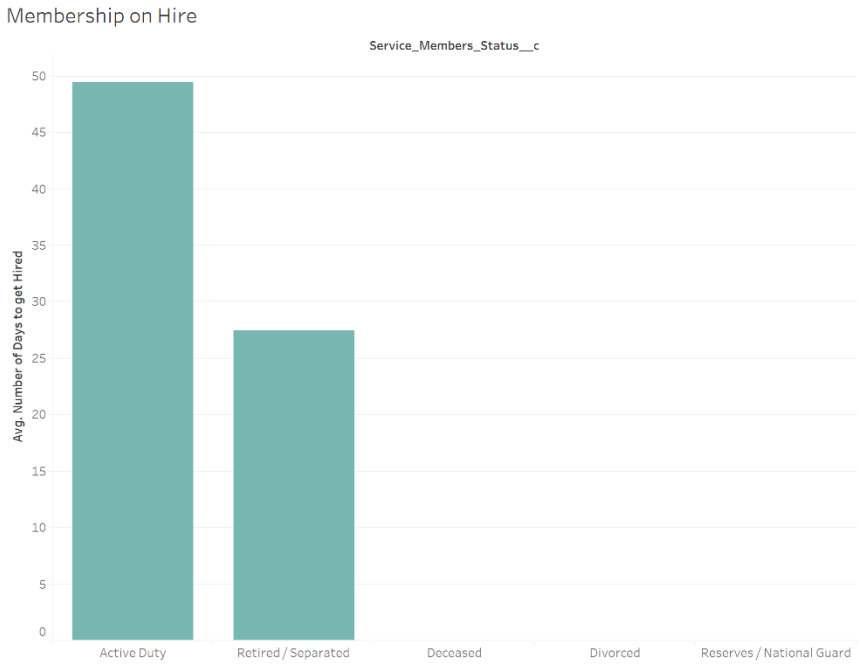




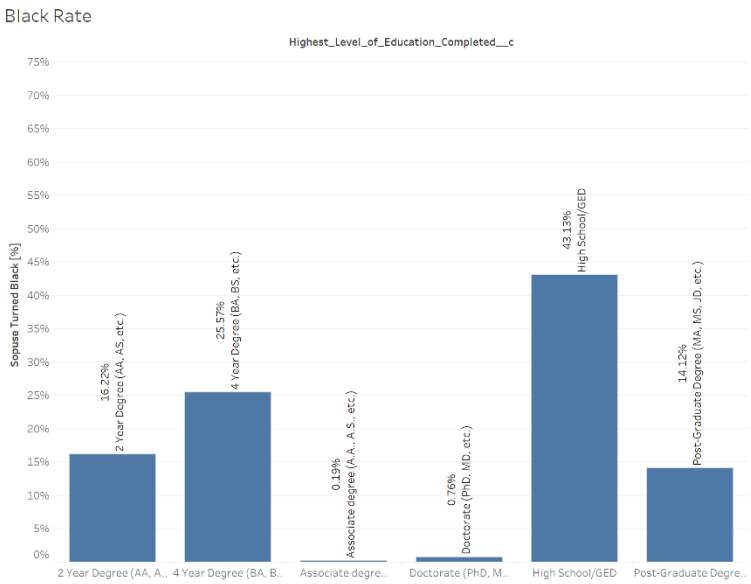
Participation of spouse state wise



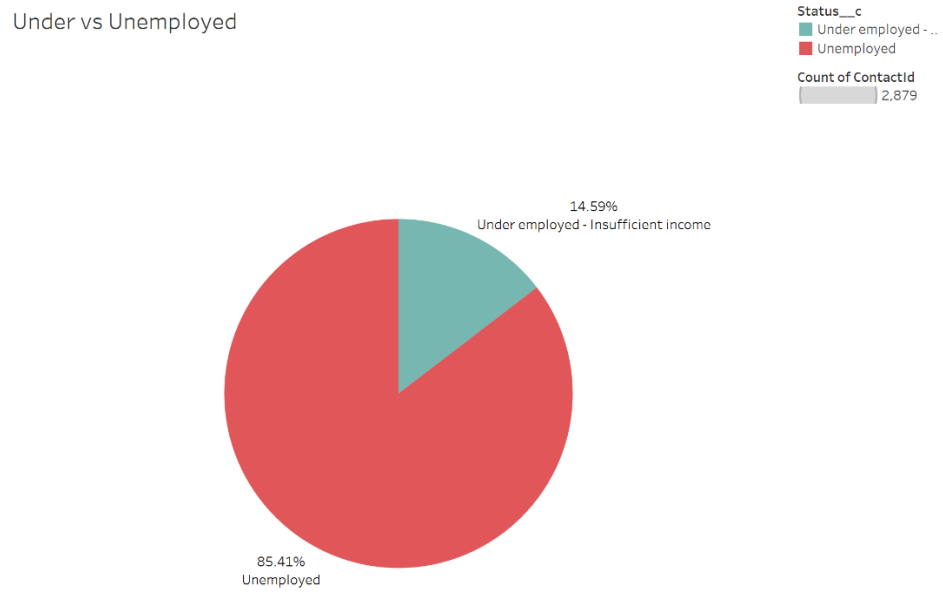
Membership status vs time to get hired



Relationship between spouse turning black and education level



percentage of military spouse clients underemployed vs. unemployed



Number of clients turn blue vs grey