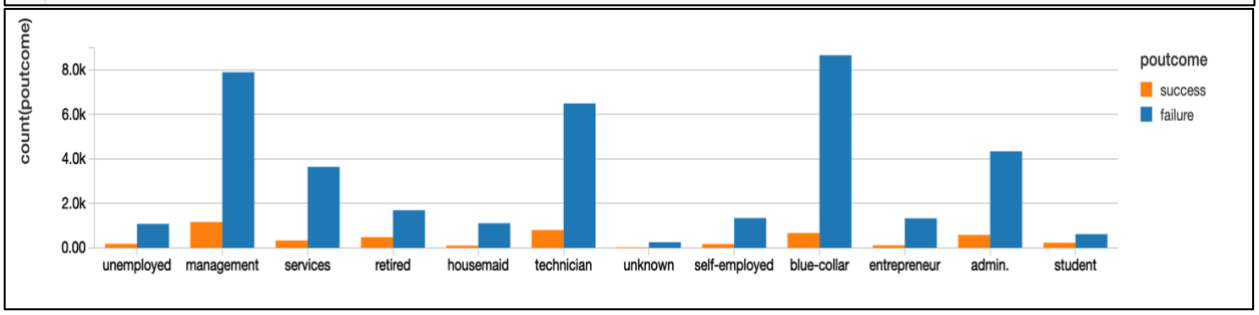
Banking Campaign Analytics

The banking data used for this coding assignment focuses on current campaign and previous campaign. The campaign is about looking for potential customers based on age, education, account balance, and marital status. Campaigning was conducted to get people invested in creating a term deposit for the current campaign. With this campaign, success and failure rates are compared to give more insights into problem areas to improve results for the current campaign and extend it to future campaigns.

Problem statement: It is very important to know the target audience to make a campaign successful. Is the campaign conducted focusing on right set of people? Has our success improved based on previous campaign? What can be done to further improve the results? All these questions can be answered based using the dataset provided.

Hypothesis 1: Based on previous and current outcomes, which group of people based on their job title have a high likelihood of saying 'Yes' to current campaign.

```
%sql
select job,count(poutcome),poutcome
from FullBankData_csv
where previous !=0 and poutcome IN ('success', 'failure')
group by poutcome,job
UNION ALL
select job,count(y),
case
when y= 'yes' then 'success'
else 'failure'
end as y
from FullBankData_csv
where previous = 0  and y IN ('yes', 'no')
group by y,job
```

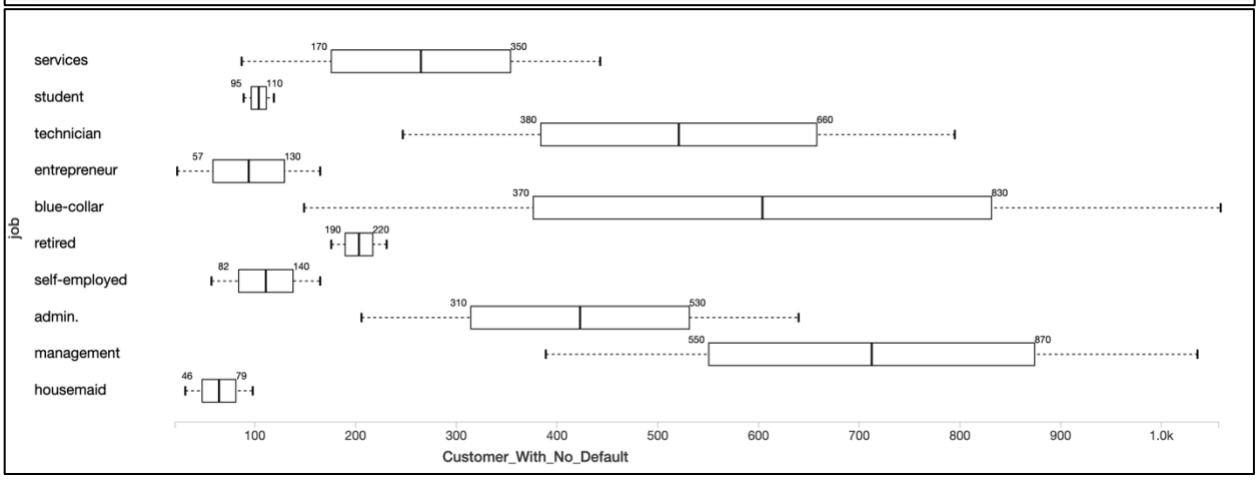


Based on the results above the following observations are made:

- The bar graph displays the cumulative success and failure of a campaigning outcome with people across having various job titles.
- The cumulative measure shows the count of total number of people with success and failure in both the previous and the current campaign.
- It is clear that the top target audience, are people who are jobs like management, services, technician, blue-collar and administration because most of the customers belong to these jobs.
- its interesting to note that, most of the customers in topmost common jobs have, with highest count, have higher failure. In the upcoming hypothesis we can examine the reason behind failure.

Hypothesis 2: Amongst top 5 job categories discovered above, finding potential customers from previous campaign who have no default(customers who have not missed any repayments to the bank in certain period of time)

```
%sql
select poutcome,default,count(default) as Customer_With_No_Default,job,
case
when default='no' and poutcome = 'success ' then 'have term deposits'
when default='yes' then 'not potential for current campaign'
else 'potential for current campaign'
end as eligibleForCurrentCampaign
from FullBankData_csv
where poutcome not in('unknown','other') and default!='yes' and job not in('unemployed','unknown')
group by poutcome,default,job
```



Based on the boxplot results, the following observations can be made:

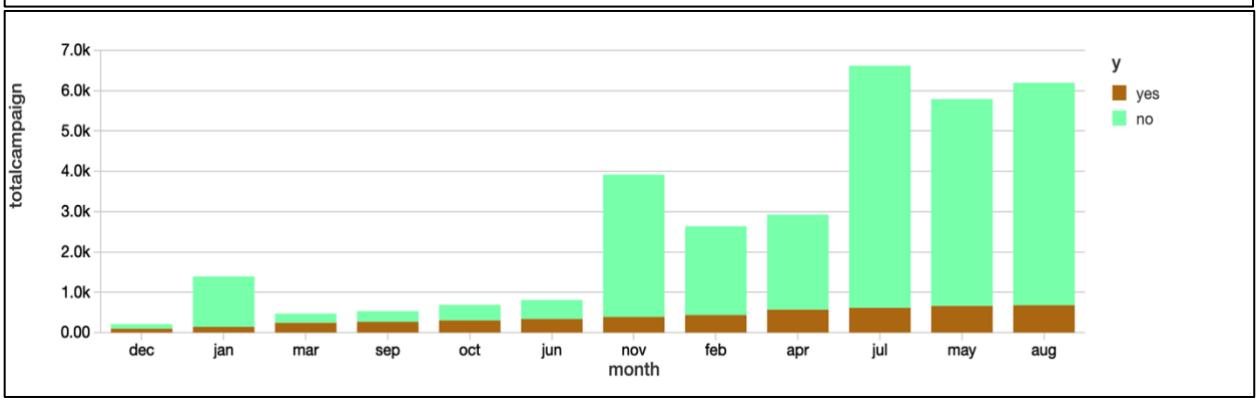
- Continuing the analysis further, based on previous campaign outcome, total number of people in management, blue-collar, technicians, administration, services are still the main target groups based on jobs.
- They are most certainly people in well paid jobs, they have no history for missing any repayments to the bank, making them potential customers for current campaign.
- Focusing on this group of people, the outcome of the current campaign can improvise.

Hypothesis 3: After finding target customers with no default history, it is important to figure out the right time to contact them to increase the success rate of the current campaign.

```

data = spark.sql("select y,month,count(campaign) as totalcampaign from FullBankData_csv where contact != 'unknown' group by y,month order by totalcampaign asc")
display(data)

```

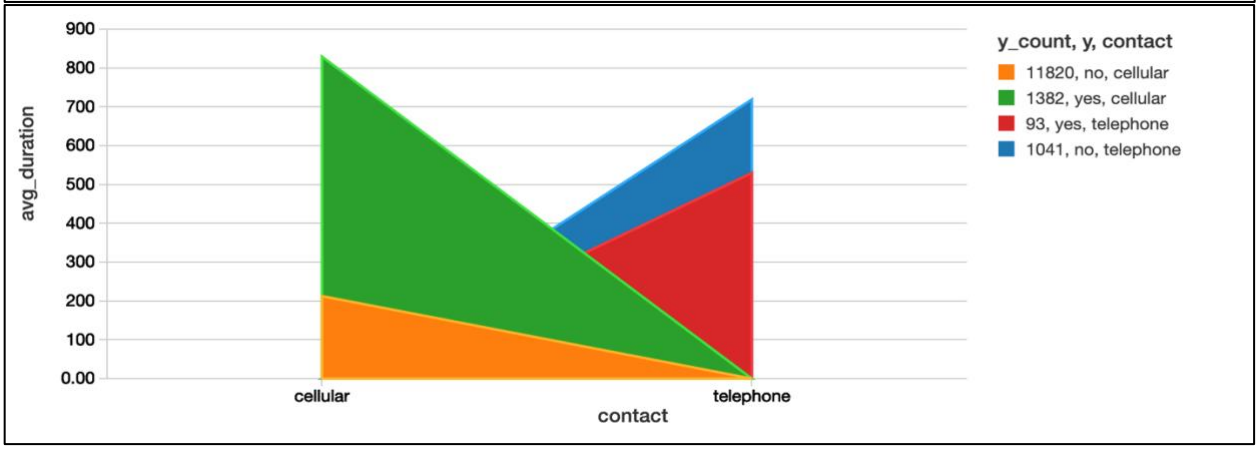


Based on the stacked bar chart above, the following observations can be made:

- The graph gives information about total number of campaigns that were conducted every month. Additionally, this also gives information about the success and failure in the current campaign.
- Looks like the best months to contact people via cellular or telephone are in months July, August and May because total number of campaigns conducted are greater than 4000.
- Not only the total number of campaigns conducted during this time is high, but also the success rate of a customer saying yes to a campaign is high.

Hypothesis 4: After finding the best months to run a campaign and the target audience in top 5 jobs, we should find out what means of communications is a better way to communicate: is it cellular or telephone? On an average, how many minutes should the customers be engaged on a call to increase the success?

```
%sql
select y, contact,
  avg(duration) as avg_duration,
  count(y) as y_count
  from FullBankData_csv
  where contact != 'unknown' and
  month in (select month from (select y,month,count(campaign) as totalcampaign
            from FullBankData_csv
            where contact != 'unknown'
            group by y, month
            having totalcampaign > 5000)) and
  job in (select job from (select job,y,count(y) as success_count
          from FullBankData_csv
          where y='yes'
          group by y,job
          having success_count > 500))
  group by y, contact
```



Based on the results above, the following observations can be made:

- The graph above is an area graph, the average duration is calculated for two modes of communication: cellular and telephone. This data is further divided to give information about the average duration for a success and failure.
- From the graph above, it is clear that maximum number of campaigns were conducted using cellular compared to telephone.
- The count of number of ‘Yes’ is significantly lower in both cellular and telephone compared to the count of number of ‘No’
- But it is interesting to know that in order to get a ‘Yes’ on cellular, average duration is greater than 800 seconds, the reason there is significantly high number if ‘No’ is the call duration was not up to the average.

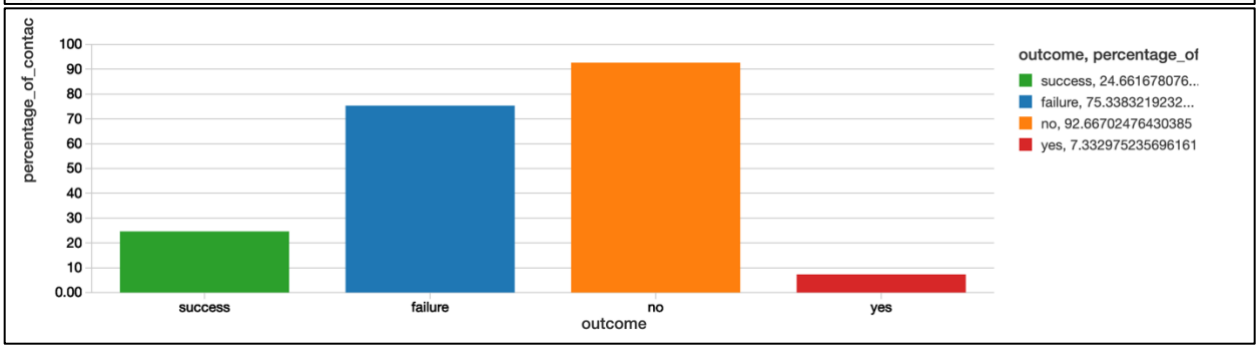
- To conclude, compared to telephone, cellular is definitely the best way to reach out to people and also by increasing the call duration greater than the average duration, the likelihood of success of high.

Hypothesis 5: after all the initial analysis, it's now time to compare success and failure of previous and current campaign? How many people were contacted in both the campaigns? Is the current campaign doing better than the previous campaign?

```
%sql

select (sum(previous)*100)/(select sum(previous) from FullBankData_csv where previous!=0 and poutcome IN ('success', 'failure')) as percentage_of_contacts,
 poutcome as outcome
  from FullBankData_csv
  where previous !=0 and poutcome IN ('success', 'failure')
  group By poutcome
UNION ALL
select (sum(campaign)*100)/(select sum(campaign) from FullBankData_csv where previous=0 and y IN ('yes', 'no')) as percentage_of_contacts, y
  from FullBankData_csv
  where previous = 0  and y IN ('yes', 'no')
  group By y

```



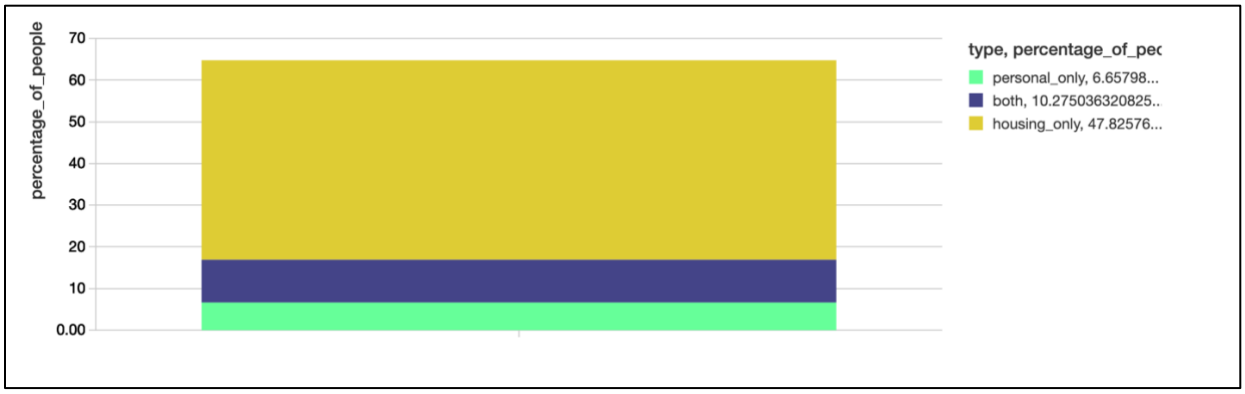
Based on the graph above, the following observations can be made:

- The graph describes outcome, where success and failure belong to previous outcome, these are distinguished from current campaign with Yes and No
- The graph gives information about the total number of people contacted for the campaign in terms of percentages.
- Clearly, with the campaigns conducted, there was high success ratio in the previous campaign compared to current campaign. The current campaign is performing significantly poor, the percentage of 'No' compared to 'Yes' is not even comparable.
- Almost 93% of the people contacted in current campaign have a 'NO', this category of NO needs a further drill down.

Hypothesis 6: After finding out the percentage of NO in current campaign, it's important to find an answer for why has the No increased from previous campaign? Is it because they have housing loans, personal loans? Or both?

```
%sql
select (count(housing)*100)/(select count(*) from FullBankData_csv where y='no') as percentage_of_people, 'housing_only' as type
  from FullBankData_csv
  where y='no' and housing = 'yes' and loan = 'no'
UNION
select (count(loan)*100)/(select count(*) from FullBankData_csv where y='no') as percentage_of_people, 'personal_only' as type
  from FullBankData_csv
  where y='no' and  loan = 'yes' and housing = 'no'
UNION
select (count(loan)*100)/(select count(*) from FullBankData_csv where y='no') as percentage_of_people, 'both' as type
  from FullBankData_csv
  where y='no' and housing = 'yes' and loan = 'yes'

```

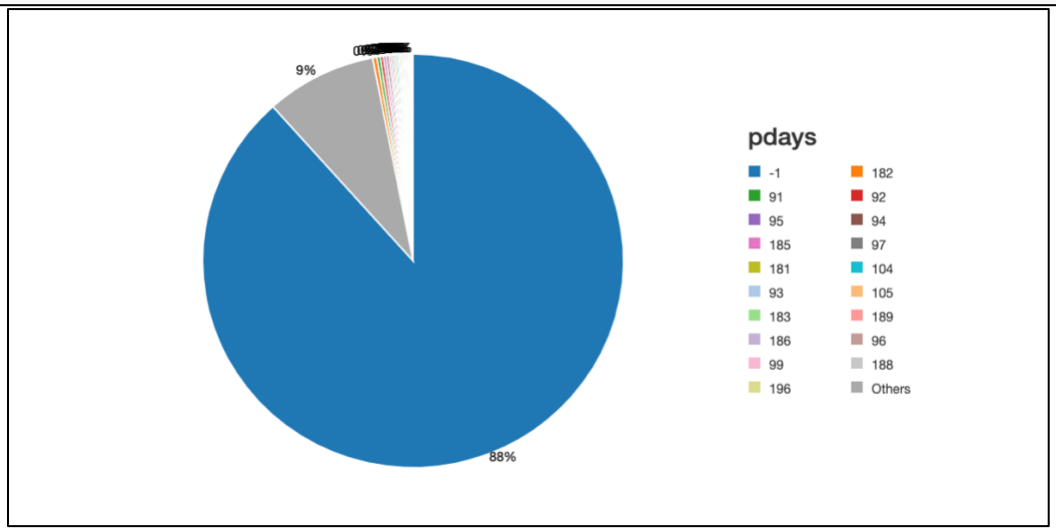


Based on the graph chart above, the following observations can be made:

- The graph above shows information of the percentage of people who have housing loans, personal loans and both who have said a 'No' to the current campaign
- It is interesting to know that, out of all the people, there are about 66 percentage of people who have loans (47%=housing loans, 6.6%=personal loans, and 10% have both housing and personal loans).
- Considering only people who have said a No to the current campaign, 66% of the people have loans, this leaves us with only 34% of the people who have agreed for the campaign on behalf of their loans.
- This gives some insights, because people with loan have to pay their monthly premium and they won't be willing to say a 'Yes' to have a term deposit.

Hypothesis 7: How about people with no loans? How many of them have said a No the current campaign? Is this current campaign overloaded with a 'No' because the customers are not frequently contacted?

```
1 from pyspark.sql.functions import explode
2
3 a=df.selectExpr('y','housing','loan','pdays').filter((df.y=='no') & (df.housing=='no') & (df.loan=='no'))
4 b=a.groupBy('pdays').count()
5 b.show
6 display(b)
```



Based on the pie chart above, the following observations can be made:

- This analysis is a continuation from the result of the above from previous hypothesis. We have 34% of the people, who have no loans of any type and yet have said No to the current

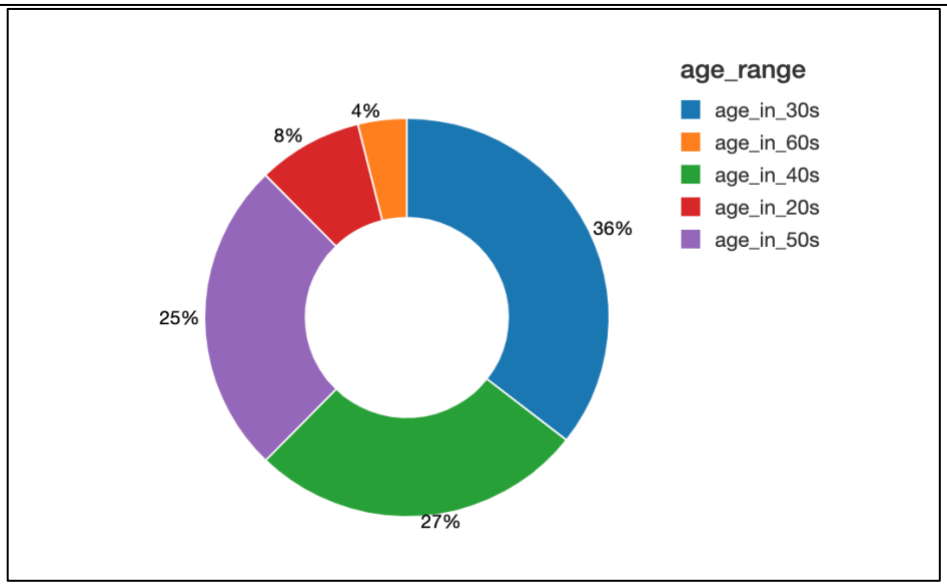
campaign. The pie chart shows information for this category of people based on the frequency of contacts made to reach out to them.

- We can infer that 88% of the people satisfying the conditions mentioned were not contacted at for the current campaign.
- In order to keep customers engaged throughout the campaign, it is important to contact them constantly, by this customer relationships can be improved. Contacting people who have no loans can increase the success of the campaign, because they might be willing to have a term deposit.

Hypothesis 8: Validating people who were left out and were not contacted are valuable customers? What age group they belong to? Will the success of the campaign increase if right age group of people are contacted?

```
%sql
Select count(*) as number_of_people, 'age_in_20s' as age_range
from FullBankData_csv
where  y='no' and housing = 'no' and loan = 'no' and pdays = -1 and
age between 20 and 29
UNION
Select count(*) as number_of_people, 'age_in_30s'  as age_range
from FullBankData_csv
where  y='no' and housing = 'no' and loan = 'no' and pdays = -1 and
age between 30 and 39
UNION
Select count(*) as number_of_people, 'age_in_40s'  as age_range
from FullBankData_csv
where  y='no' and housing = 'no' and loan = 'no' and pdays = -1 and
age between 40 and 49
UNION
Select count(*) as number_of_people, 'age_in_50s'  as age_range
from FullBankData_csv
where  y='no' and housing = 'no' and loan = 'no' and pdays = -1 and
age between 50 and 59
UNION
Select count(*) as number_of_people, 'age_in_60s' as age_range
from FullBankData_csv
where  y='no' and housing = 'no' and loan = 'no' and pdays = -1 and
age between 60 and 69

```



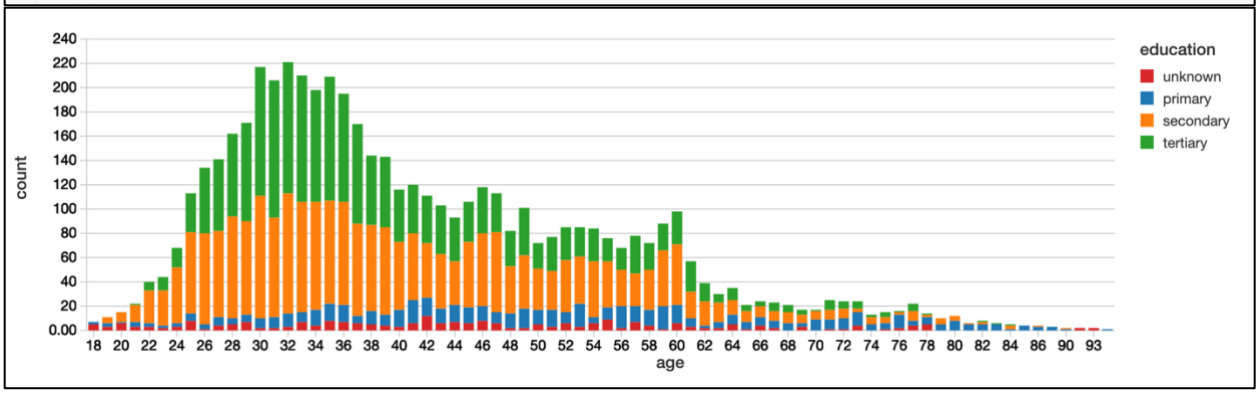
Based on the donut pie chart, the following observations can be made:

- The chart gives the percentage of people in different age intervals who were not contacted for the current campaign. They are also people with no loans but have been categorized with people who have a ‘No’ as outcome in current campaign.
- It can be observed that the maximum number of people are in their 20s,30s, 40s and 50s.
- The likelihood of getting customers interested in terms deposit is high amongst these age groups because they would have stable jobs and will be willing to explore new options to save money.

Hypothesis 9: Does contacting people in mentioned age groups impact on success of the outcome? How is the distribution of people based on age who have said ‘Yes’ to current campaign? Can the current campaign be made more specific to people who are in certain age groups with certain education levels?

```
from pyspark.sql.functions import explode

a=df.selectExpr('education','y','housing','loan','pdays','age','marital','balance','job').filter((a.y=='yes'))
b=a.groupBy('age','education').count().sort('age')
b.show
display(b)
```



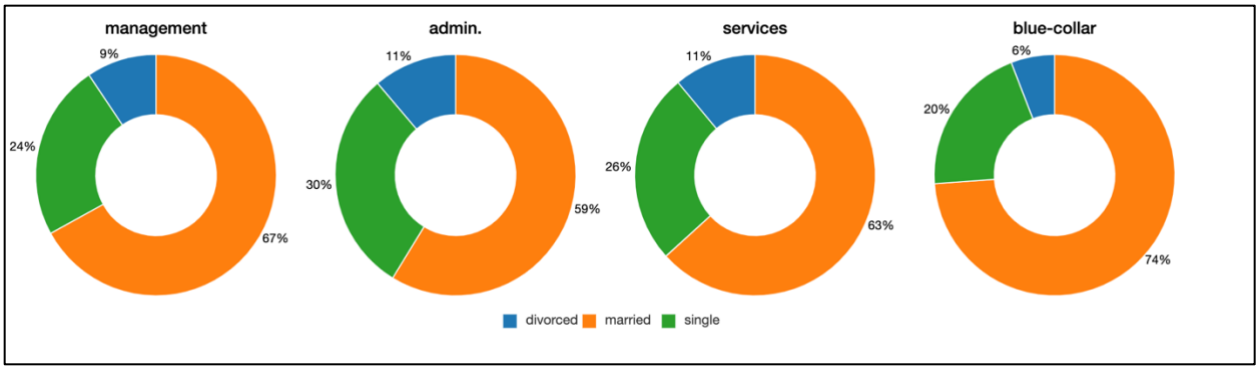
After analyzing the data for people with ‘No’ as outcome in current campaign, we can shift out analysis to people who have said ‘Yes’ to current campaign. Based on the above stacked bar chart, the following observations can be made:

- Looks like just like people with ‘No’ as outcome who have not been contacted, the people with ‘Yes ’ as outcome are also in the age groups 20s,30s, 40s and 50s. Taking this analysis further, it can be made sure that potential customers who are left out can increase the success of the current campaign.
- Also, people in these groups, most of them have secondary and tertiary as their education levels.
- To connect this analysis to hypothesis 8, we should contact the target people as early as possible to increase the success rate.

Hypothesis 10: Extending the analysis from previous hypothesis, what is the balance of the people in the targeted age groups? Do people in topmost common job categories have high balance? Are people who are married have high balances?

```
from pyspark.sql.functions import explode

a=df.selectExpr('education','y','housing','loan','pdays','age','marital','balance','job').filter((a.y=='no') & (a.housing=='no') & (a.loan=='no') & (a.pdays==-1)
& (a.age>30) & (a.age < 60) & (a.education!= 'unknown') & (a.education!='primary') & (a.job=='management') | (a.job=='technician') | (a.job=='admin.') |
(a.job=='services') | (a.job=='blue-collar'))
b=a.groupBy('marital','job').agg({'balance':'sum'})
b.show
display(b)
```



After analyzing all the hypothesis and taking, the following results and charts are derived. The final hypothesis focuses on the right people who need to be contacted as soon as possible to increase the success. The following observations can be made:

- This hypothesis is extension of the 9th hypothesis. And there are 4 pie charts representing top 4 jobs with highest number of people. These charts are further grouped based on marital status.
- It is clear that majority of the portion belongs to people who are married followed by people who are single. These two categories of people have higher balance across top 4 job categories like management, administration, services and blue-collar.

Executive summary: After all the hypothesis conducted, the performance of the current campaign is significantly low compared to the previous campaign. To improve the results the following actions needs to be taken to make this campaign a success.

- Contact people who have not been contacted after the previous campaign and have been registered as people who are not interested in the current campaign.
- Contact people using cellular mode of communication during May, July and August because wider audience can be reached during this time.
- Focus more on people:
    - In jobs like management, blue-collar, administration, services, technician
    - In the age interval from 20 – 60
    - With education level as secondary and tertiary
    - With marital status as married and single

  These following target customers are most likely having higher balances and would most likely agree for having term deposit. These people will definitely increase the success rate.

Additional analysis in Tableau:

The following visualization was created using Tableau, the visualization is a dashboard. This was created to provide more analysis and support the executive summary. The following analysis are made:

- The first horizontal bar chart, Job vs Balance, gives information about the category of jobs who have sum of balance greater than average, and they are management, blue-collar, administration and technician, services and retired.
- The second line graph, Current campaign vs Day and Job, provides information about total number of campaigns conducted based on date. This by default considers all the months. It is interesting to notice that more campaigns are conducted after 15th of each month with 20th being the highest. It further analyzes number of people contacted in various jobs mentioned in first chart.
- The third scatter plot, Age vs Balance, provides information about success and failure for current campaign considering how age and balance influence the outcome. With the trend

lines, for both success and failure they have similar trend. But we can infer that most people are in age interval 20-60 and they have balance anywhere between $-1500 - $20,000

- The fourth scatter plot, Balance vs age, provides information about success and failure for current campaign considering how duration and balance influence the outcome. With the trend lines different from each other for 'Yes' and 'No' it can be inferred that people saying 'No' to the current campaign has a little or no influence on duration and balance. But, for people who have said 'yes' to the campaign it is interesting to notice that as the balance increased the duration decreases, indicating that they are people with good balances and are not interested to have a term deposit.