Data analytics report

Linh Tran

Data at a first glance

Data component: 2 data sets, *'accounts'* and *'account_metrics'*

Number of attributes:

- 'accounts': 9 variables
- 'account_metrics': 38 variables

Number of records:

- 'accounts': 9,546
- 'account_metrics': 1,357

Data period:

- 'accounts': from 12th of Mar 2020 to 22th of Jan 2021
- 'account_metrics': data with latest service created from 18th of Feb 2016 to 21th of Jan 2021

Data Quality Assessment: ‘accounts’ dataset

➢**Accuracy**: there are 3,938 accounts created on 12th of Mar 2020 in only 1 hour from 14.00 to 14.59 => there may be an issue of accuracy in this part of the data (these observations may be correct but the time they were created does not seem right)

➢**Completeness:** no values in ‘continent’ and ‘country’ columns

‘main_industry’ column has 9,066 missing values

‘sub_industry’ column has 9,513 missing values

‘support_tier’ has 9,503 missing value

‘type’ has 1 missing value

➢**Consistency:** the values in dataset are consistent

➢**Validity: t**he values in dataset are valid

➢**Uniqueness:** there is no duplicate records in the dataset

➢**Currency:** the dataset is updated until 22th of Jan 2021 => up-to-date

Data Quality Assessment: ‘account_metrics’ dataset

➢**Accuracy**: 2 outlier observations (index: 36,63) in which 1 outlier (index 36) has inaccurate values in the observation

There are 45 rows with inaccurate values of active services or active projects

➢**Completeness:** the data set is quite completed

➢**Consistency:** the values in ‘cloud_regions_in_use_names’ (after parsing JSON to normal text) are not consistent, different words were used to name a same region.

➢**Validity:** the values in ‘cloud_regions_in_use_names’ and ‘cloud_in_use_names’ are JSON data that needed some work to simplify it. Also the JSON key '{"v":]}' in both columns are incorrect JSON key for NA value.

➢**Uniqueness:** there is no duplicate records in the dataset

➢**Currency:** the dataset is updated until 21th of Jan 2021 => up-to-date

Mitigation:

‘accounts’ dataset:

➢ Exclude 3,935 records of accounts created on 12th of Mar 2020 when starting analytics based on date time for better results. Since the assumption is that these observation are valid but their created time is not => include them in analytics that is not date time analytics.

➢ Delete columns: ‘continent’, ‘country’, ‘main_industry’, ‘sub_industry’

‘account_metrics’ dataset:

➢ Exclude 45 rows with inaccurate values of active services or active projects

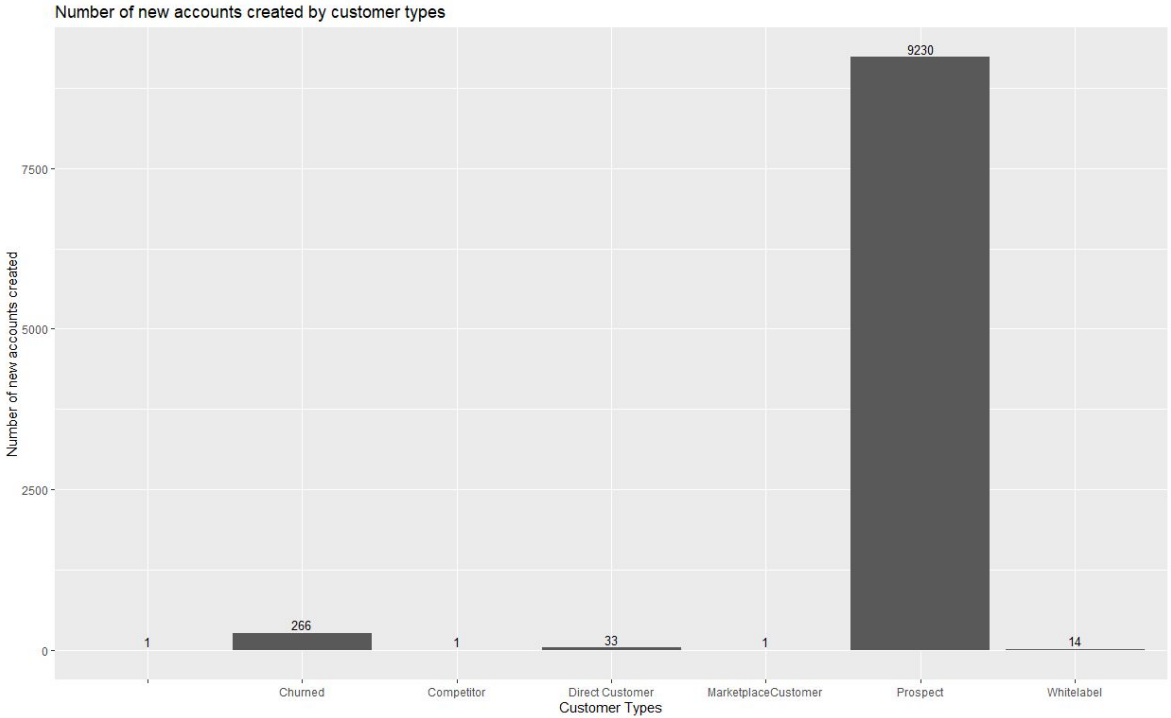
➢ Replace JSON key '{"v":]}' by '{"v":[{"v":"NA"}]}'

➢ Group the cloud regions into 3 main group: America, Asia Pacific, and Europe

EDA: ‘accounts’

➢ There is 1 observation missing customer type value

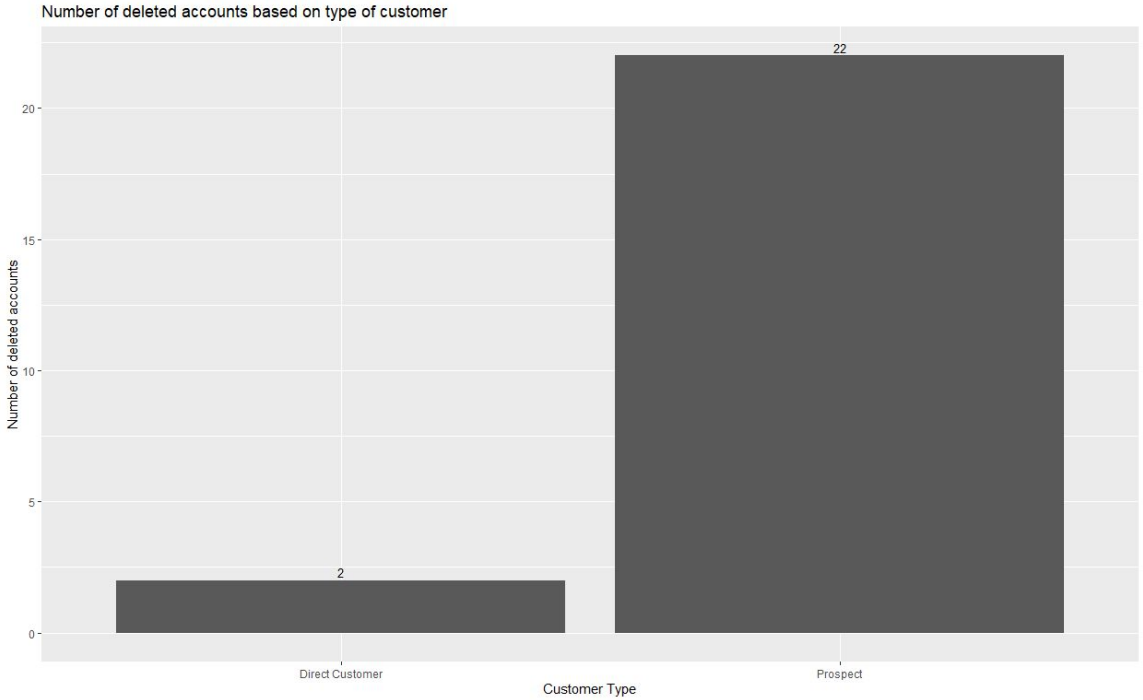
➢ The majority of new accounts created during data period is Prospect customer



EDA: ‘accounts’

➢ There are 24 accounts deleted in the data period.

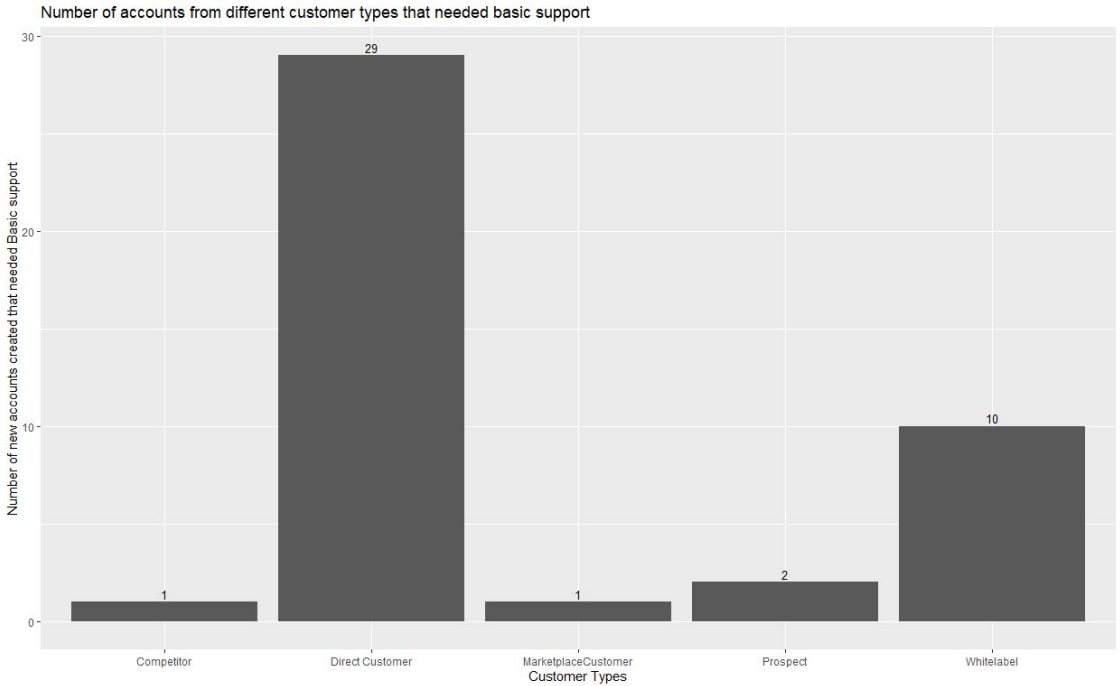
➢ Most of the accounts deleted are of the Prospect customer type.



EDA: ‘accounts’

➢ 26 of 30 new accounts created of Direct Customer needed basic support

➢ 10 out of 14 new Whitelabel accounts created needed basic support.



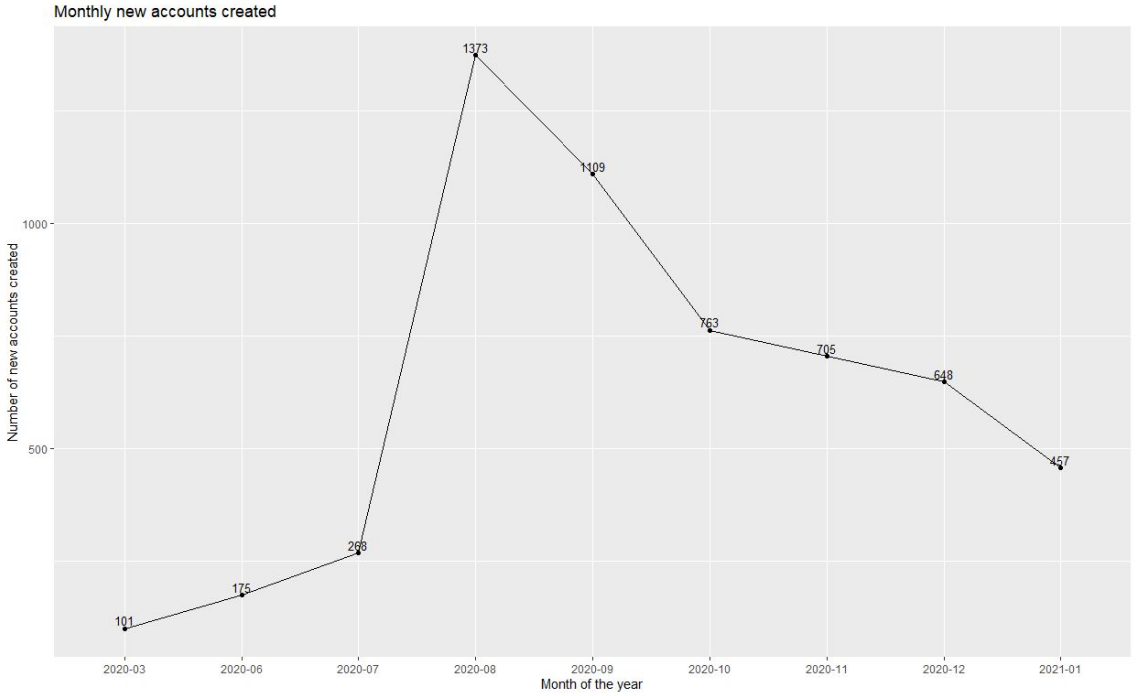
EDA: ‘accounts’

➢ Deleted 3,935 records of accounts created on 12th of Mar 2020 for better time line data analytics

➢ No new account was created in April and May 2020

➢ A big jump up in the number of new accounts created during August 2020

➢ After August 2020, the number of new accounts created decreased steadily

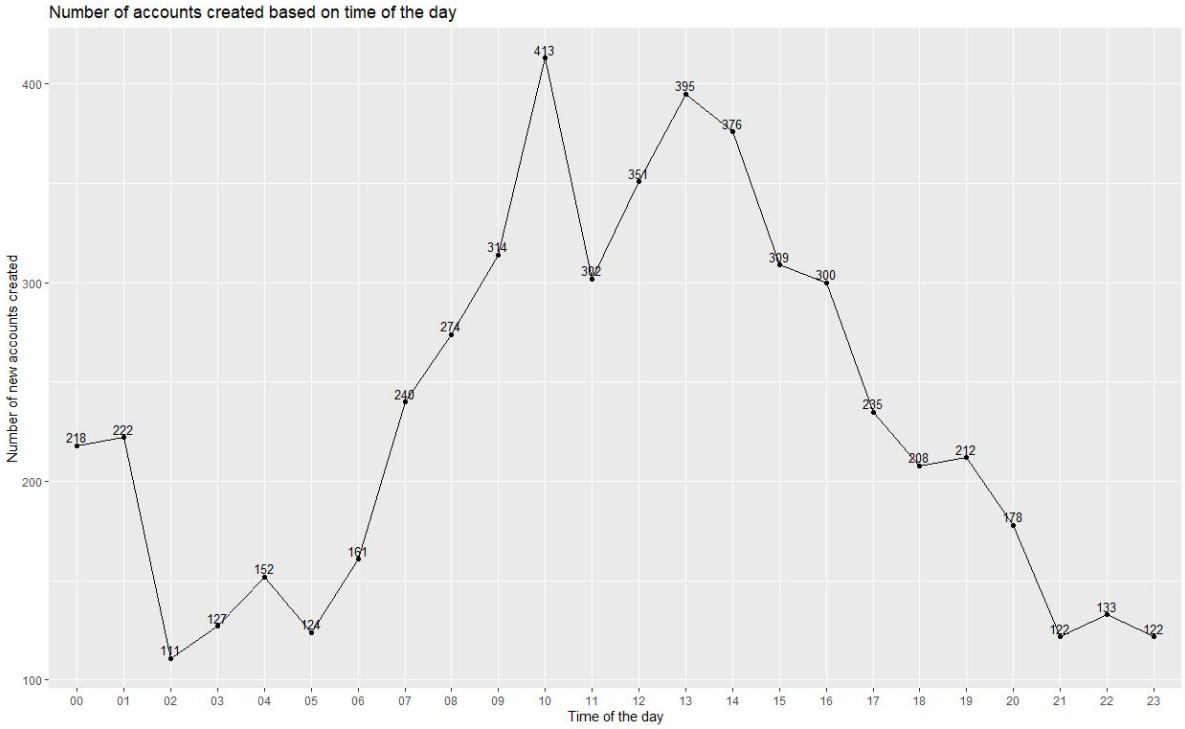


EDA: ‘accounts’

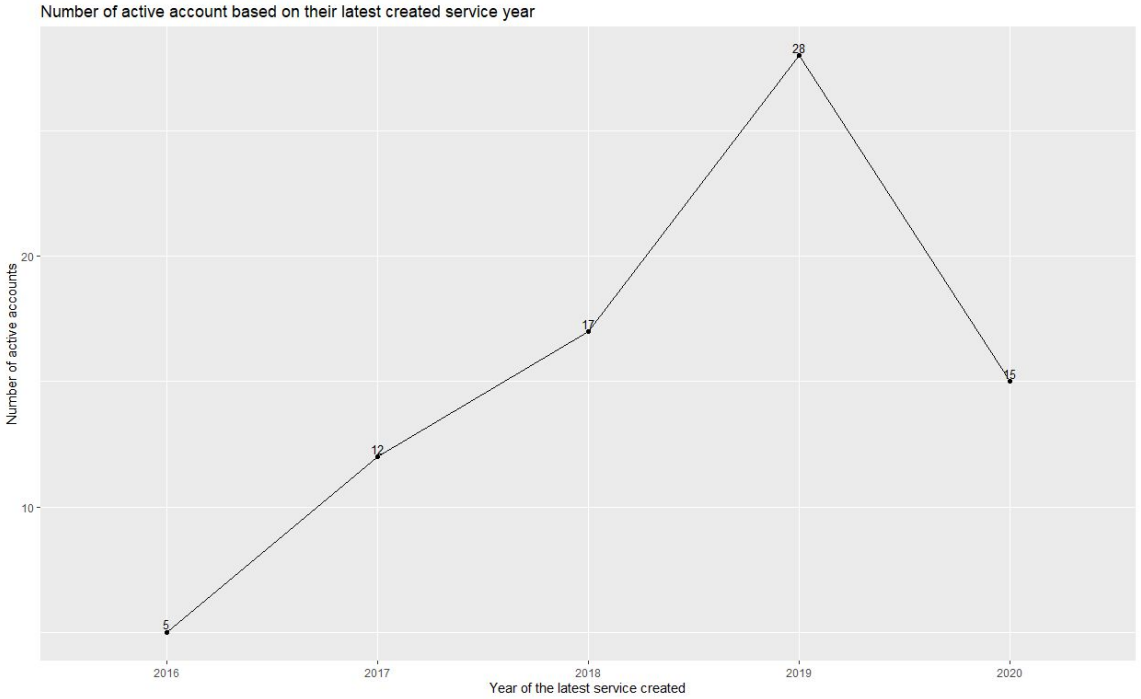
➢ Deleted 3,935 records of accounts created on 12th of Mar 2020 for better time line data analytics

➢ More accounts were created during the time from 8 a.m to 4 p.m (during working hour)

➢ There also a considerable number of accounts created during midnight.



EDA: ‘account_metrics’

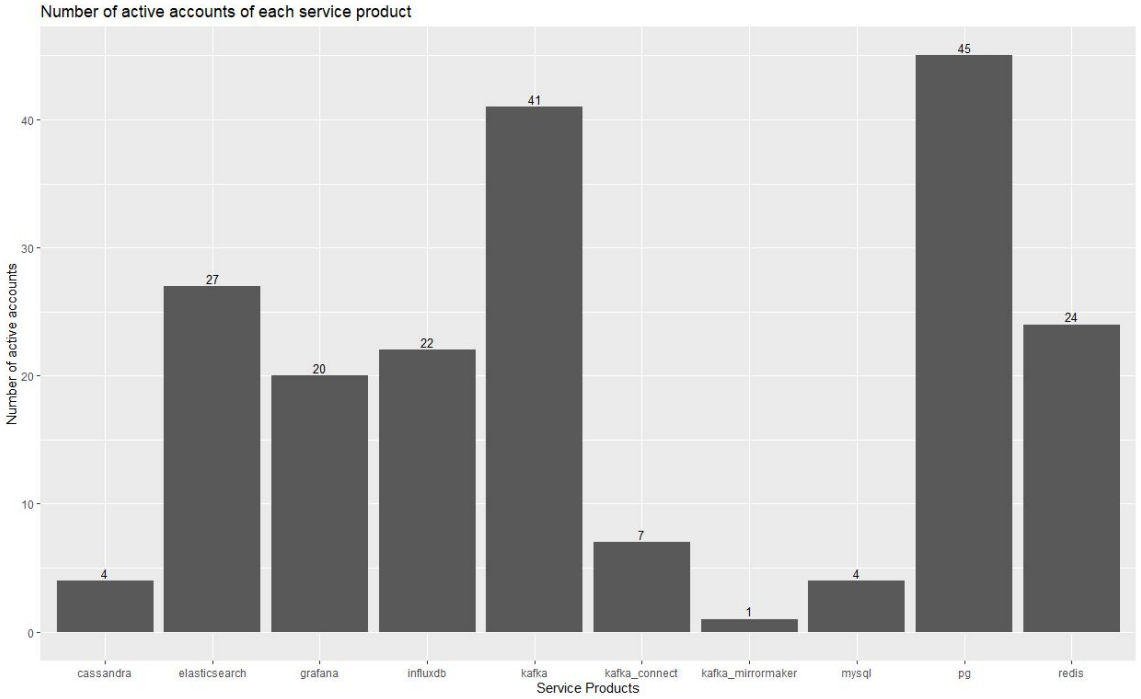


➢ After deleted the inaccurate observations: there are 1,312 distinct accounts in the dataset

➢ Out of 1,312 accounts, there are only 77 accounts which have active projects or services

➢ The more recent service created, the more active accounts (expect for the accounts with latest service created in 2020)

EDA: ‘account_metrics’



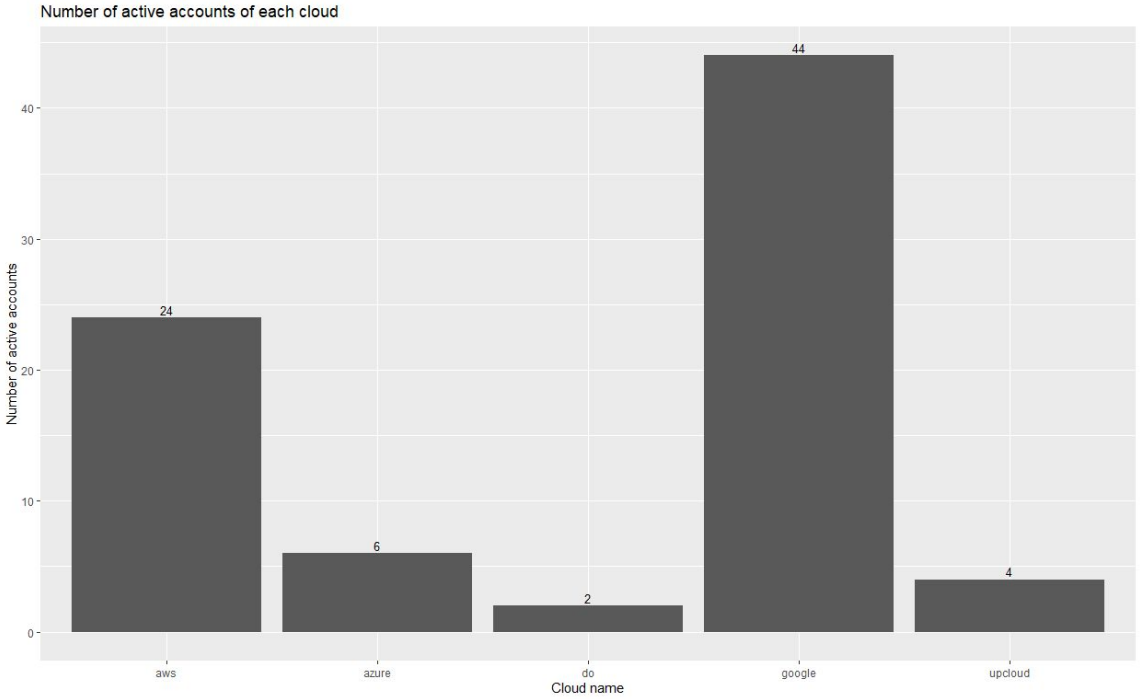
➢ There are total of 12 service products

➢ Kafka and PG are services with highest number using accounts.

➢ Elasticsearch, Grafana, Influxdb, Redis services follow with similar using accounts

➢ Cassandra, Kafka connect, Kafka mirrormaker, My SQL services have relatively low amount of accounts using.

EDA: ‘account_metrics’

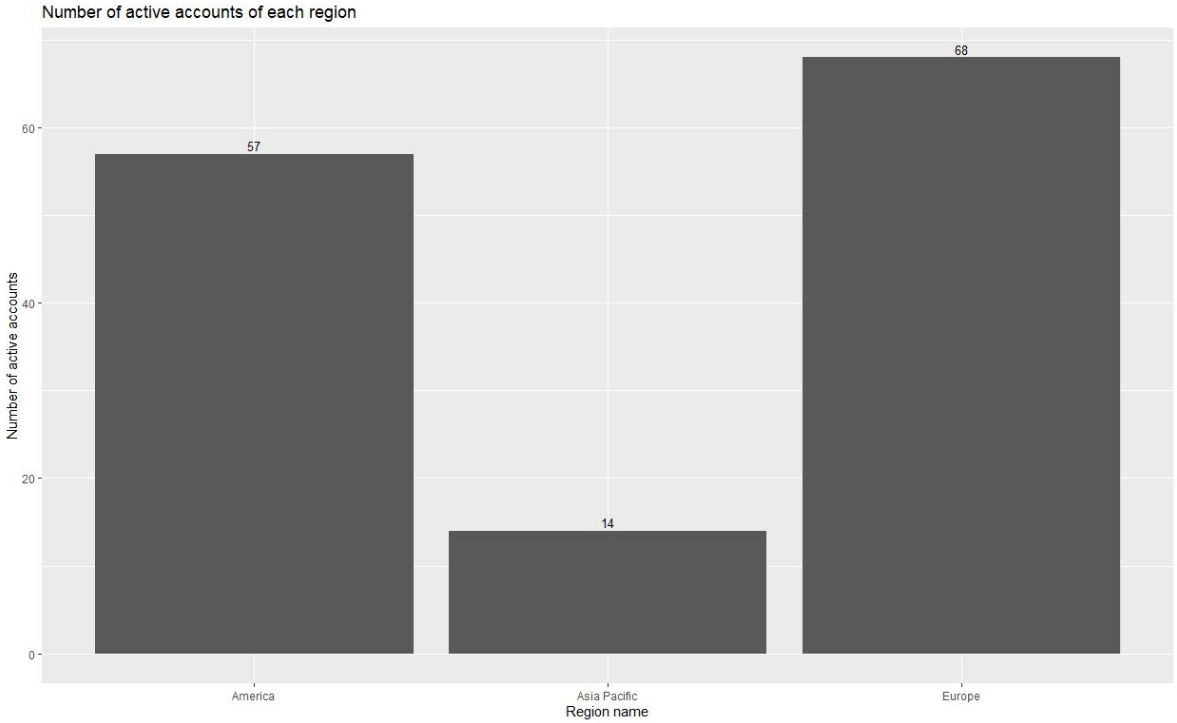


➢ Google cloud has the highest amount of user accounts.

➢ DO cloud doesn’t do really well with only 2 accounts using this cloud.

➢ No accounts used Custom DO and Packet cloud products

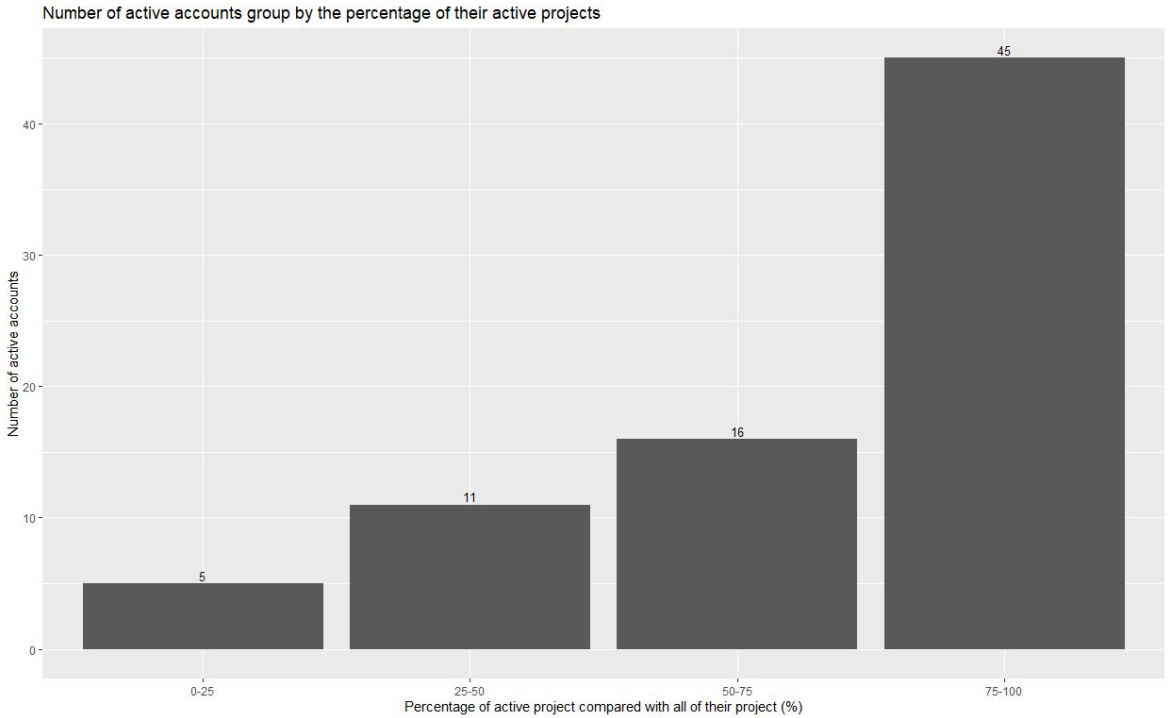
EDA: ‘account_metrics’



➢ Most of active accounts comes from Europe and America.

➢ The number of active accounts in Asia Pacific is relatively low

EDA: ‘account_metrics’

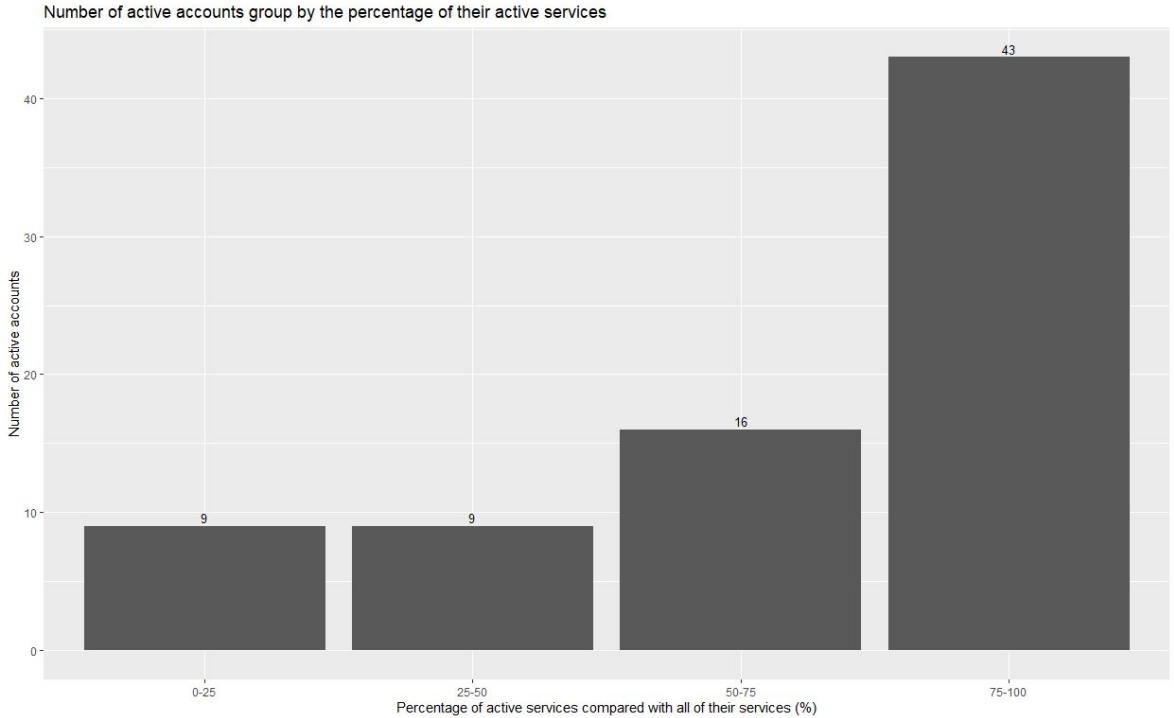


➢ There are 77 out of 1,312 accounts have projects created in their portfolios

➢ Out of 77 accounts created projects, 45 of them have the percentage of active projects higher than 75%

=> most of users, if they have created their projects, they tend to have their projects active.

EDA: ‘account_metrics’



➢ The amount of account in each group of percentage of active services spreads similarly with the active projects

➢ A small difference: the amount of account have 0-25% of their services active is a bit higher than account have 0-25% of their projects active (9 compare with 5 accounts)

‘accounts’ vs. ‘account_metrics’: Inner Join

Inner Join:

➢ Keep all observation of ‘accounts’ table for inner join and exclude the 45 rows in ‘account_metrics’ table with inaccurate values of active services or active projects

=> There are 498 records in this inner join table

➢ When analyze the latest_service_created time and account_created_time, also exclude 3,935 records of accounts created on 12th of Mar 2020 for better time line data analytics

=> There are 171 records in this inner join table

➢ There is no account in inner join table that has active projects or active services

➢ There are 30 observations out of 171 that have latest service created time after the account created time

➢ There are 141 observations out of 171 that have latest service created time before the account created time

➢ Users can create service before creating account?

Conclusion (1):

Data quality of 2 datasets is not ideal, some cleaning needed to perform in order to get the data in better shape.

2 datasets are not consistent with each other, merging 2 datasets doesn't have much results.

Should users create account first then create the services? Is there any accuracy issues in term of latest_service_created date and account_created_time?

Conclusion (2):

'accounts'

➢ Most of new accounts created are Prospect customers

➢ 24 accounts deleted during the data period, most of them are Prospect customers

➢ Most of the customers that needed basic support are Direct and White Label customers

➢ There was no new account created in April, May 2020. Is this an accuracy problem of the data or something happened in April, May that prevent users from creating new account?

➢ There was a huge increase in the number of new account created in August, 2020. Is this a seasonal trend or what happened that encourage this increase?

'account_metrics'

➢ Assumptions: the number of active projects should be equal to sum of each project, sum of active services and stopped services should be equal to all services, active services should be equal to the sum of all service products => there are 45 observations that don't fit with these assumptions => excluded

➢ Only 77 accounts that currently have active projects or services

➢Kafka and PG are the services that have the highest amount of active accounts

➢ In term of cloud, Google and AWS have the highest amount of active accounts

➢ America and Europe are the regions that have the highest amount of active clouds

➢ Most of accounts that are active, they have 75% to 100% of all of their projects or services active.

Thank you!