Project 3 - The Most Valued Data Science Skills

Kossi Akplaka, Carol Campbell, Saloua Daouki, and Souleymane Doumbia

2023-10-29

Using data to answer the question…

“Which are the most valued data science skills?”

Our Methodology

- Collaborative tools
  - Created dedicated GitHub repository for all code
  - Slack
    - Quick messaging; scheduling meetups
  - Zoom
    - Held several meetings to discuss progress
    - Share code, real time revisions
- Use R Studio to clean and visualize our data
  - Tidyverse, stringr and gridExtra libraries of choice
- Search web for suitable data set
  - Kaggle.com
    - “Data Analyst Skill Analysis” by Luke Barousse

Tidying the data

Steps we took…

1. Study raw data
2. Determine what we need for analysis
3. Plan analysis
4. Code
5. Test
6. Repeat steps 4 & 5
7. Present findings

The beginning:

Our dataset that contained 43 columns and 15 variables – too many for our analysis, so we used various methods to clean the data

Filter data set into 3 separate careers…

- use grepl function to create new titles based on description

Getting 3 separate careers

```{r}
skills <- df %>%  mutate(new_title = case_when(
grepl("ETL|pipeline|warehouse|architect", description, ignore.case = TRUE) ~
"Data Engineer",    grepl("machine learning | modeling|A/B testing", description,
ignore.case = TRUE) ~ "Data Science",    TRUE ~ "Data Analyst"   )) tibble (skills)
```

A tibble: 1,820 x 44

X <int>	index <int>	title <chr>
30	30	Data Engineer/Data Analyst
31	31	Healthcare Data Analyst
32	32	Remote Training & placement Data Analyst
33	33	Clinical Data Analyst - Queen's Clinically Integrated Physician...
34	34	Product Information Management (PIM) Data Analyst
35	35	Financial Data Analyst
36	36	Provider Maintenance Data Analyst
37	37	Data Analyst, Education & Services
38	38	Sr. Healthcare & Accounting Data Modeler and Analyst
39	39	Analyst III - REMOTE

31-40 of 1,820 rows | 1-3 of 44 columns

Data Analyst	Data Engineer	Data Science
1012	481	327

Tidying before…

A tibble: 1,820 x 8

skills <chr>
[]
['sql', 'excel', 'python']
['r', 'sql', 'azure', 'python']
['power_bi']
['sql', 'jira']
['python', 'tableau', 'r', 'sql', 'excel']
['power_bi', 'ssis', 'python', 'tableau', 'dax', 'powershell', 'r', 'git', 'c', 'sql', 'ssrs', 'excel', 'azure']
['go', 'spark', 'c', 'sql']
['python', 'excel']
['spss', 'tableau', 'excel']

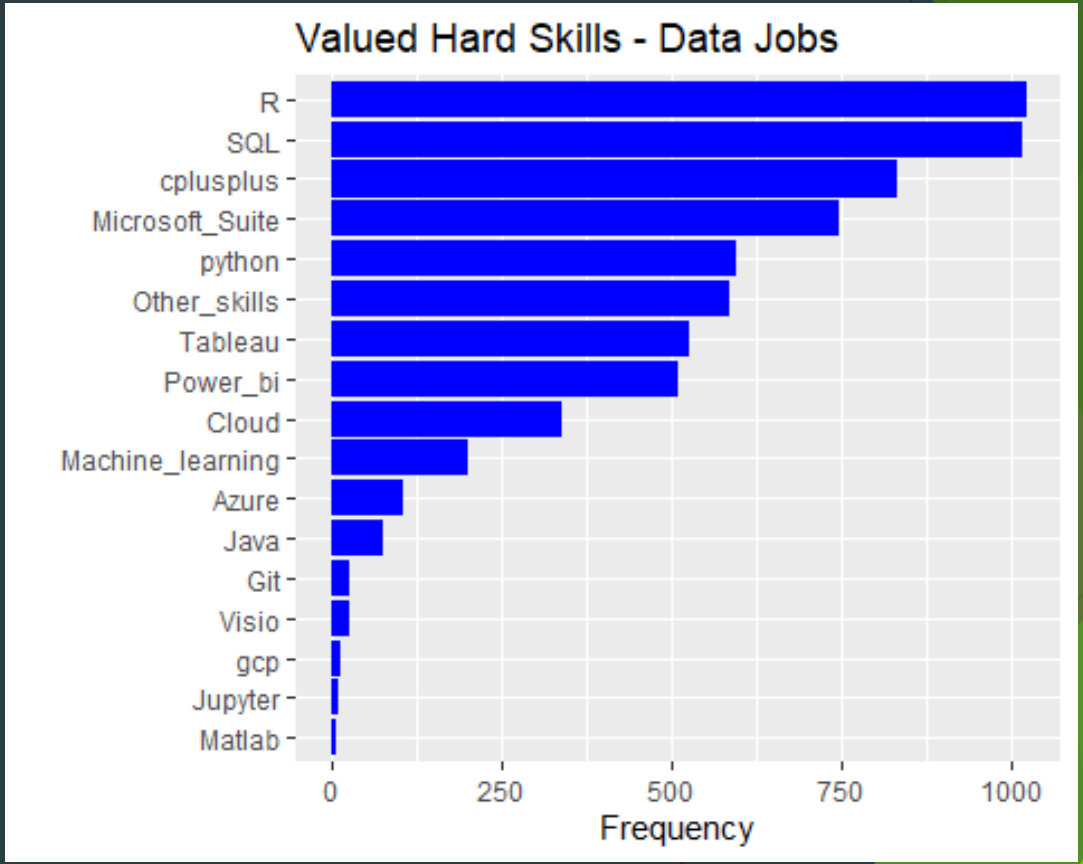
and after…

A tibble: 1,820 x 1

skills <chr>
sql excel python
r sql azure python
power bi
sql jira
python tableau r sql excel
power bi ssis python tableau dax powershell r git c sql ssrs excel azure
go spark c sql
python excel
spss tableau excel

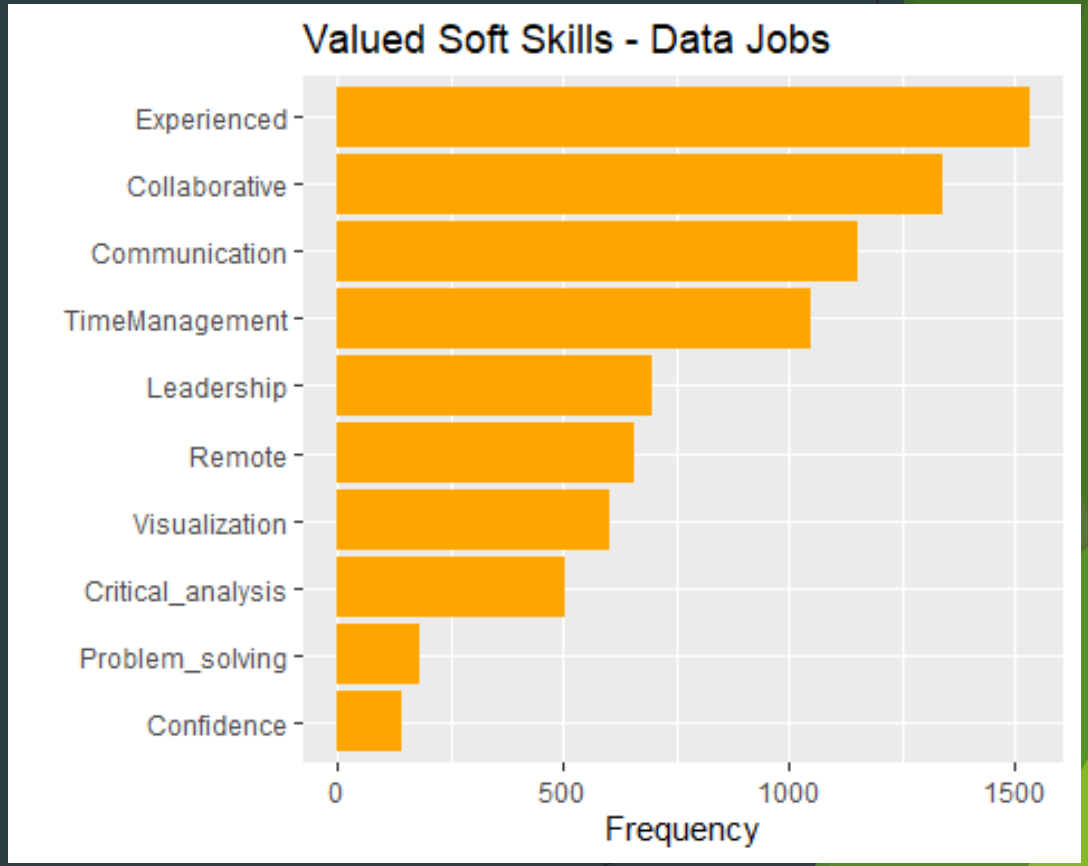
Hard skills for all Data Jobs

```{r hard skill plot}

ggplot(hardskills,aes(x=reorder(skill, freq), y=freq)) +
geom_bar(stat='identity',fill="blue") + xlab('') + ylab('Frequency') +
labs(title='Valued Hard Skills - Data Jobs') + coord_flip()
```



Soft skills for all Data Jobs

```
ggplot(softskills,aes(x=reorder(skill, freq), y=freq)) +
geom_bar(stat='identity',fill="orange") +
xlab('') + ylab('Frequency') +
labs(title='Valued Soft Skills - Data Jobs') +
coord_flip()
```



CONCLUSION
FOR ALL DATA JOBS

Based on our analysis, we concluded that the top 5 valued hard skills for ALL Data Jobs

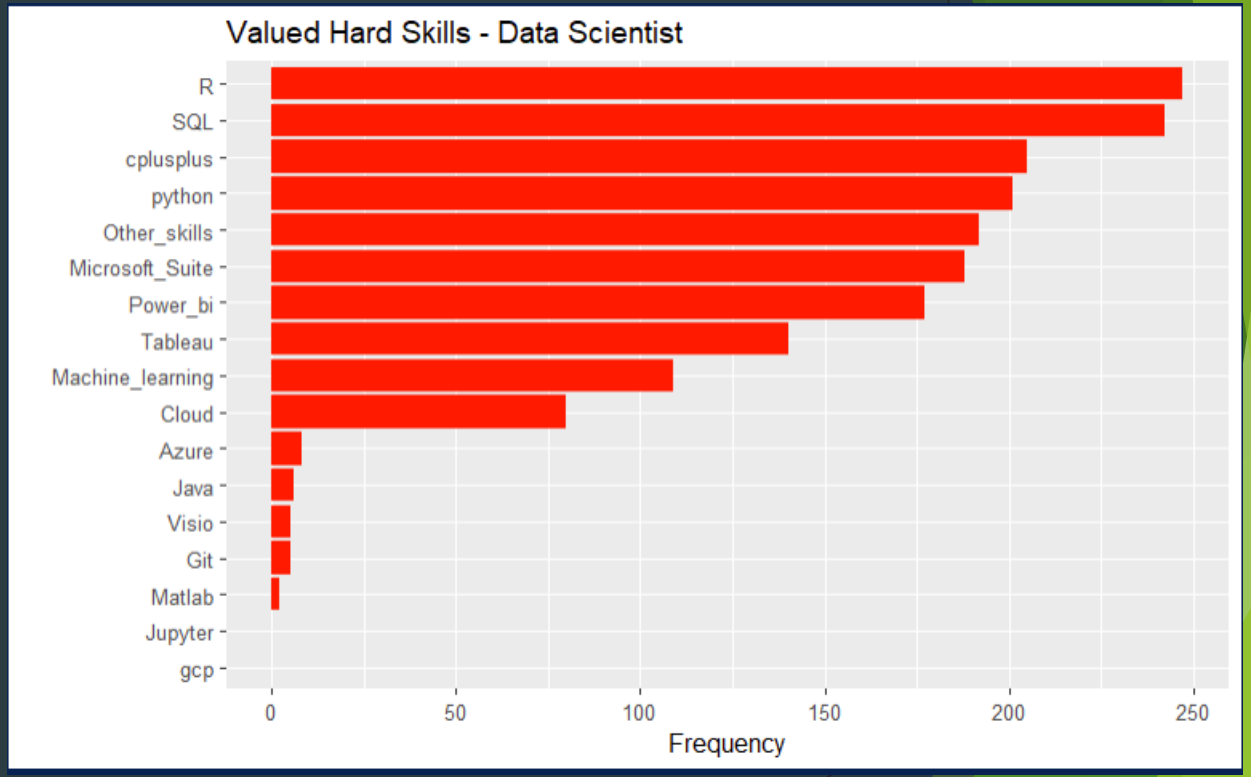
- R
- SQL
- C++
- Microsoft Suite
- Python

Similarly, the top 5 valued soft skills are:

- Experience
- Collaboration
- Communication
- Time management
- Leadership

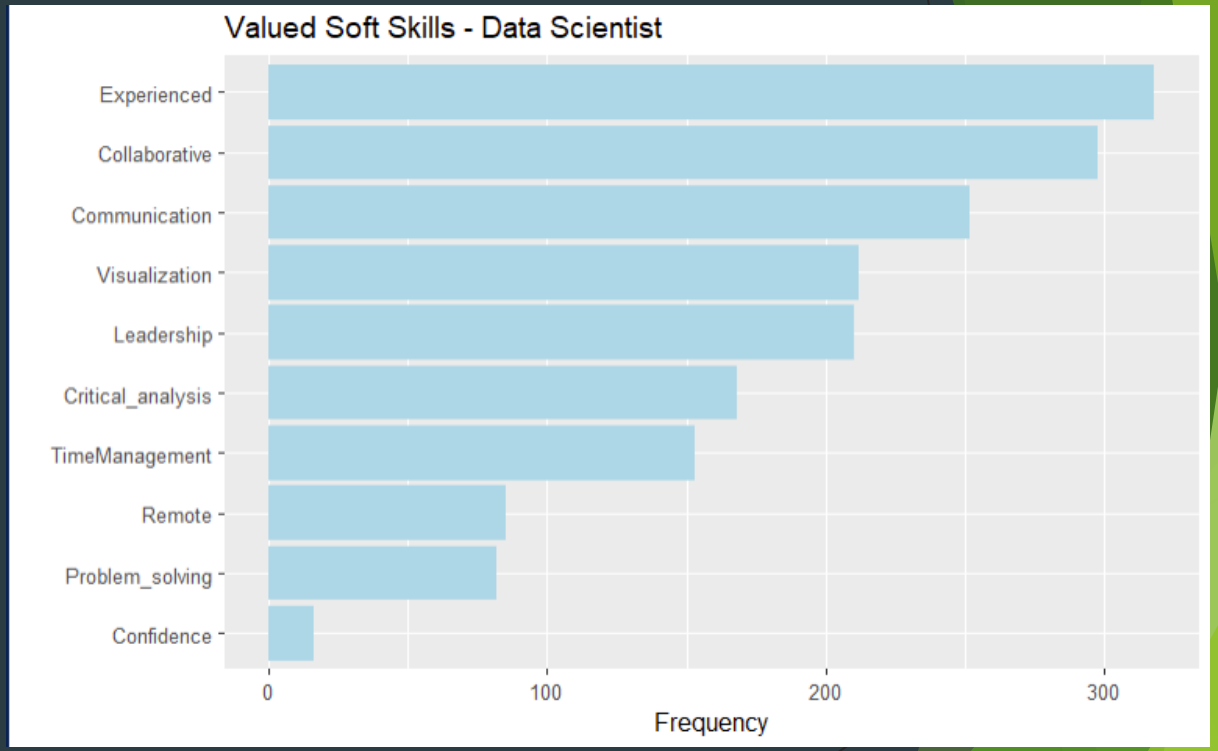
Hard skills for Data Scientist

```{r plotting the hard skills for data science}

ggplot(hardskills_science,aes(x =reorder(skill, freq), y=freq)) + geom_bar(stat='identity',fill="red") + xlab('') + ylab('Frequency') + labs(title='Valued Hard Skills - Data Scientist') + coord_flip()

```



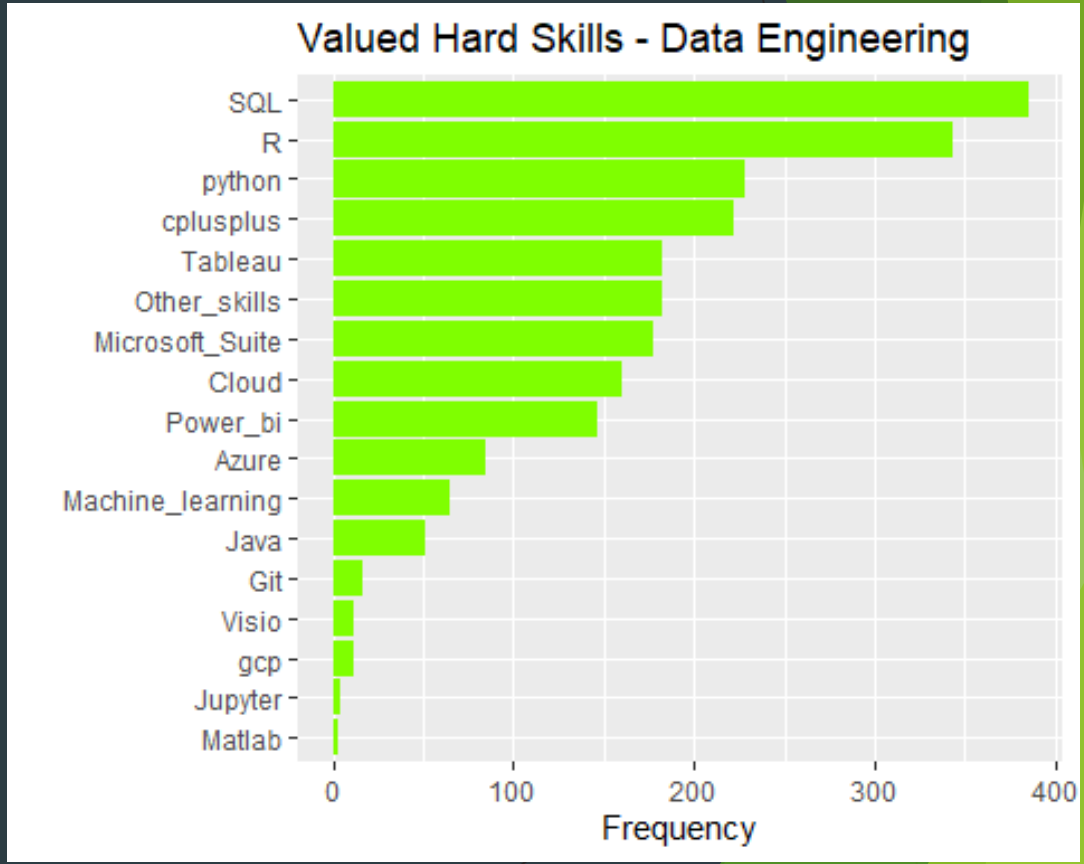
Soft skills for Data Scientist

```{r plotting the soft skills for data science}

ggplot(softskills_science,aes(x =reorder(skill, freq), y=freq)) +
geom_bar(stat='identity',fill="lightblue") + xlab('') +
ylab('Frequency') +
labs(title='Valued Soft Skills - Data Scientist ') +

 coord_flip()

```



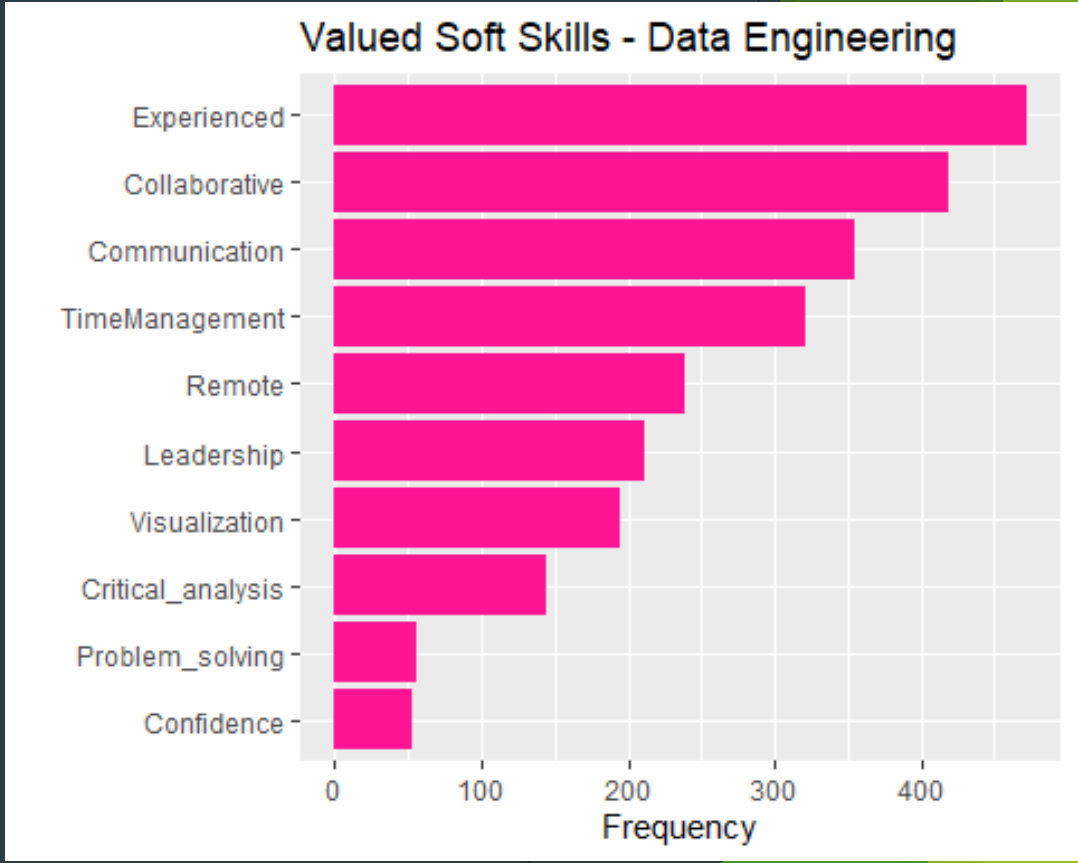
Hard Skills Data Engineering Jobs

```{r plotting the hard skills for data engineer}

ggplot(hardskills_engineer,aes(x=reorder(skill, freq), y=freq)) + geom_bar(stat='identity',fill="chartreuse1") + xlab('') + ylab('Frequency') + labs(title='Valued Hard Skills - Data Engineering') + coord_flip()

```



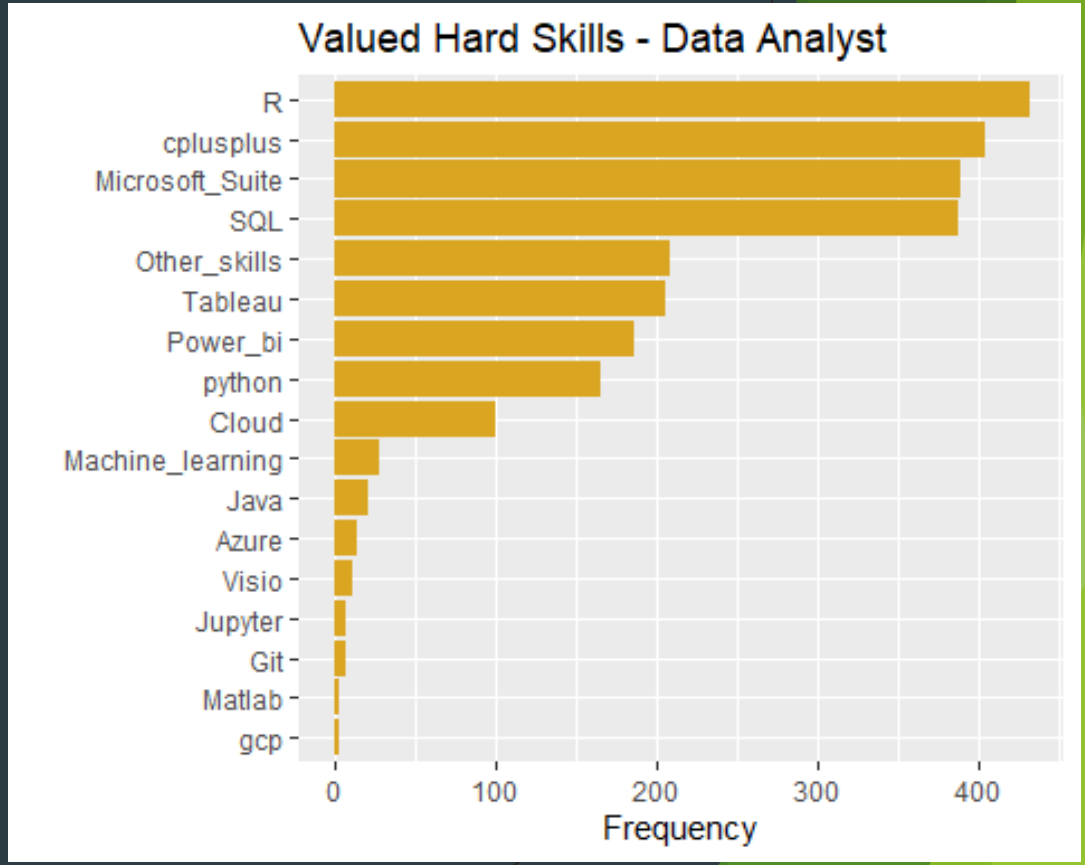
Soft Skills Data Engineering Jobs

```{r plotting the soft skills for data engineer}

ggplot(softskills_engineer,aes(x= reorder(skill, freq), y=freq)) + geom_bar(stat='identity',fill="deeppink") + xlab('') + ylab('Frequency') + labs(title='Valued Soft Skills - Data Engineering') + coord_flip()
```



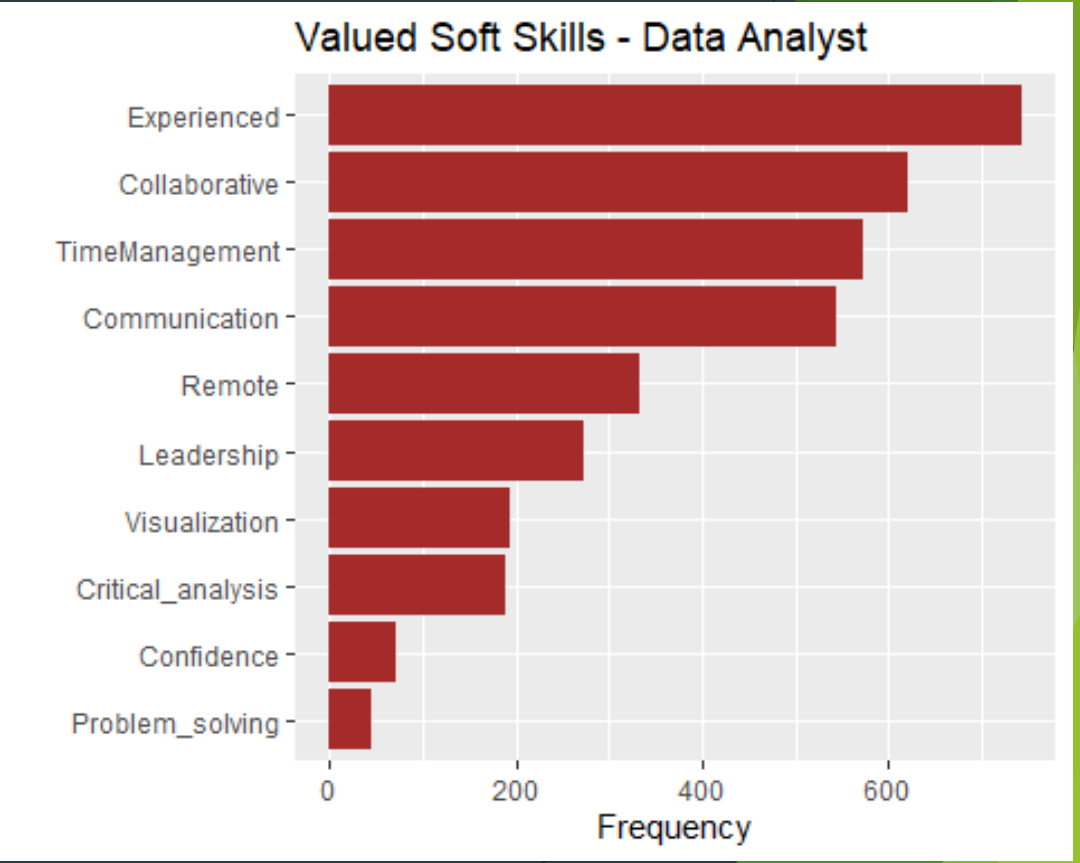
Hard Skills Data Analyst Jobs

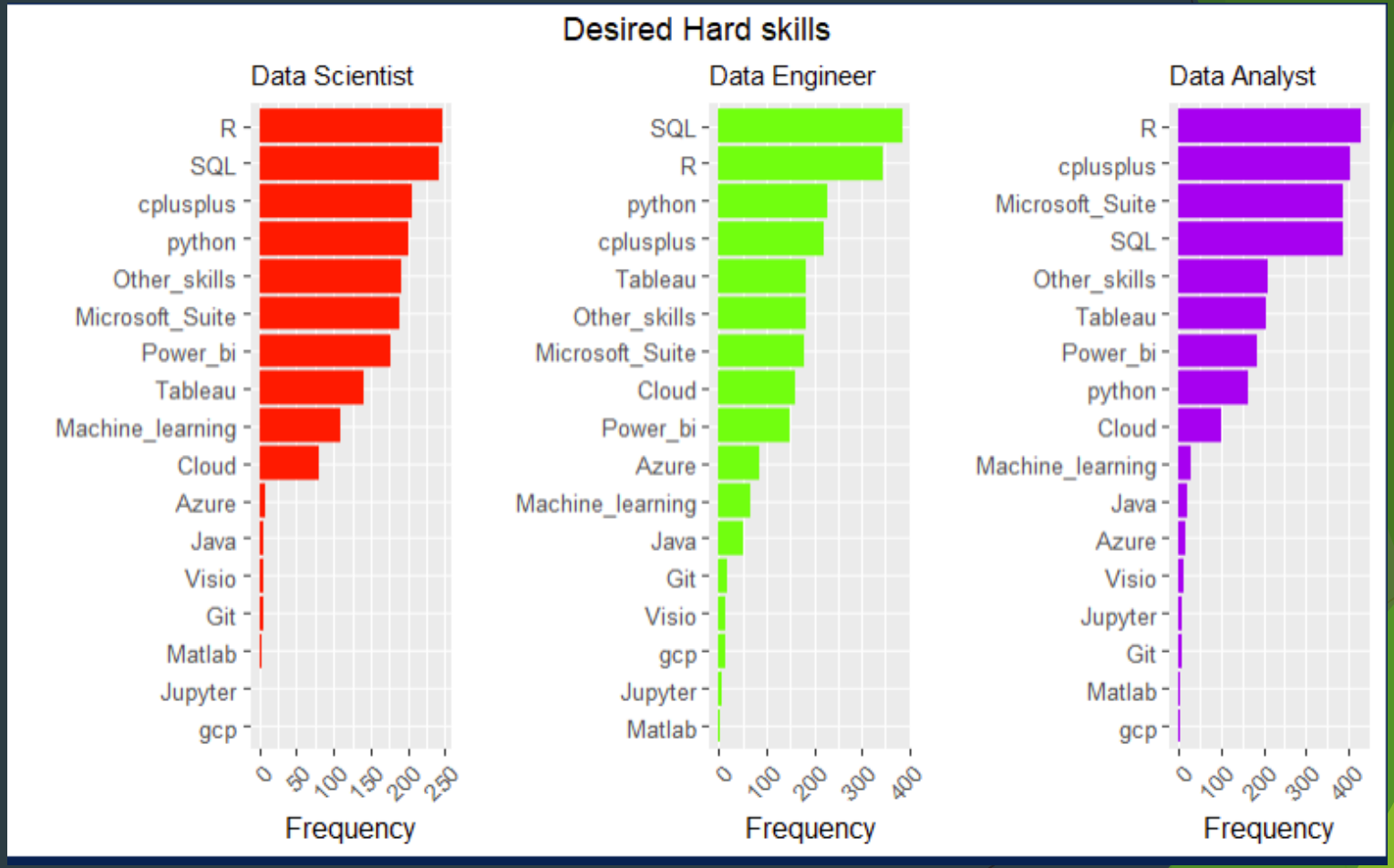
```{r plotting the hard skills for data analyst}

ggplot(hardskills_analyst,aes(x=reorder(skill, freq), y=freq)) + geom_bar(stat='identity',fill="goldenrod") + xlab('') + ylab('Frequency') + labs(title='Valued Hard Skills - Data Analyst') + coord_flip()

```

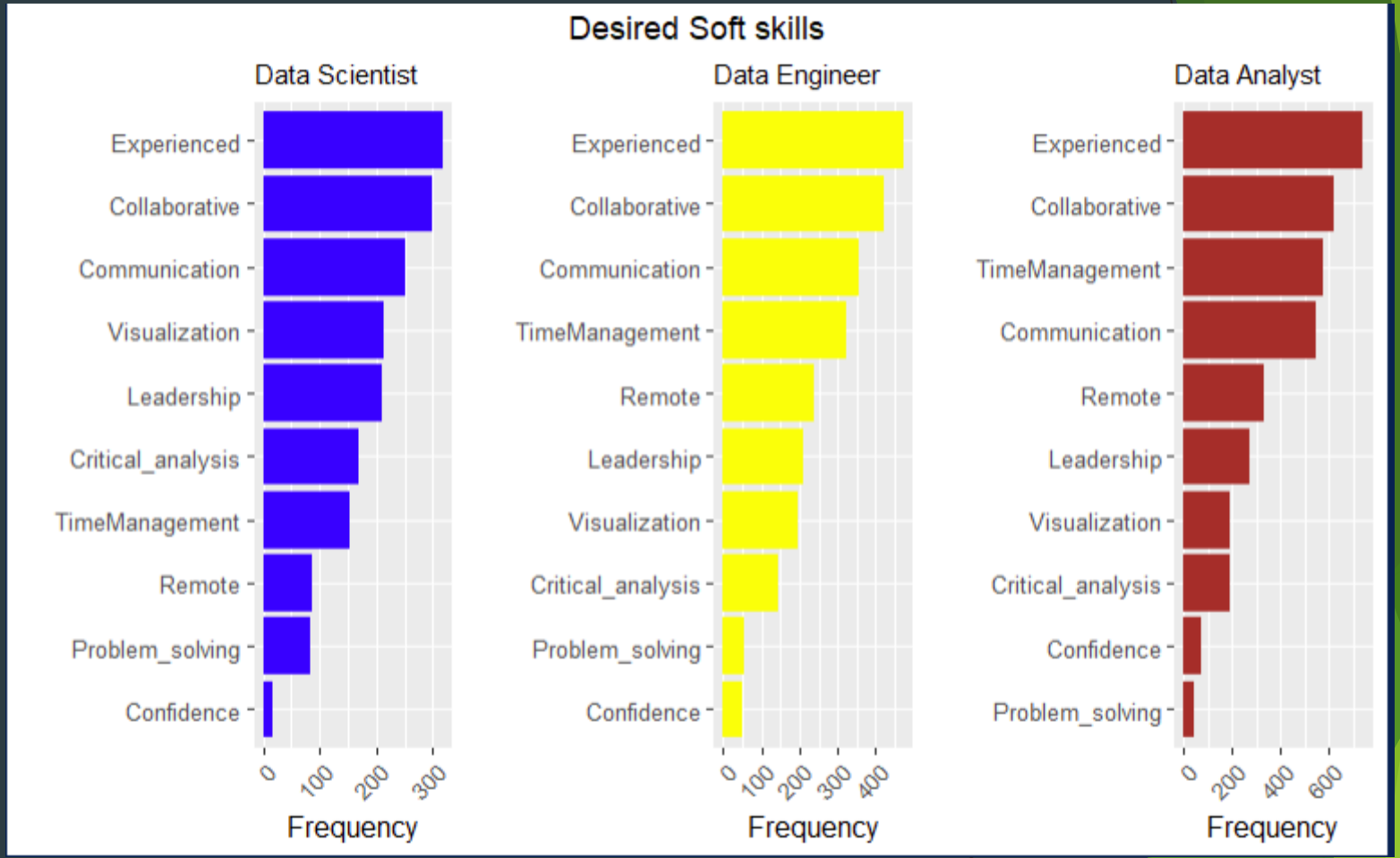


Soft Skills Data Analyst Jobs

```{r plotting the soft skills for data analyst}

ggplot(softskills_analyst,aes(x=reorder(skill, freq), y=freq)) + geom_bar(stat='identity',fill="brown") + xlab('') + ylab('Frequency') + labs(title='Valued Soft Skills - Data Analyst') + coord_flip()

```




Desired Hard skills
Data Scientist
R
SQL
cplusplus
python
Other_skills
Microsoft_Suite
Power_bi
Tableau
Machine_learning
Cloud
Azure
Java
Visio
Git
Matlab
Jupyter
gcp
0
50
100
150
200
250
Frequency
Data Engineer
SQL
R
python
cplusplus
Tableau
Other_skills
Microsoft_Suite
Cloud
Power_bi
Azure
Machine_learning
Java
Git
Visio
gcp
Jupyter
Matlab
0
100
200
300
400
Frequency
Data Analyst
R
cplusplus
Microsoft_Suite
SQL
Other_skills
Tableau
Power_bi
python
Cloud
Machine_learning
Java
Azure
Visio
Jupyter
Git
Matlab
gcp
0
100
200
300
400
Frequency


Desired Soft skills
Data Scientist
Experienced
Collaborative
Communication
Visualization
Leadership
Critical_analysis
TimeManagement
Remote
Problem_solving
Confidence
0
100
200
300
Frequency
Data Engineer
Experienced
Collaborative
Communication
TimeManagement
Remote
Leadership
Visualization
Critical_analysis
Problem_solving
Confidence
0
100
200
300
400
Frequency
Data Analyst
Experienced
Collaborative
TimeManagement
Communication
Remote
Leadership
Visualization
Critical_analysis
Confidence
Problem_solving
0
200
400
600
Frequency

CONCLUSION

The plots above indicate that the most 5 most valued soft skills for the following jobs are:

- **Data science:** experience, collaboration, communication, visualization , leadership.
- **Data engineer:** experience, collaboration, communication, time management , remote.
- **Data analyst:** experience, collaboration, time management, communication , remote.