Decent work in the Digital Era : Data Science Trends

1. 2702244284 - Renata Aqila Ridha Putri 2. 2702247683 - Cresenshia Hillary Benida 3. 2702222163 - Sherly Vaneza 4. 2702246945 - Sophiana Kassandria Sukamto

2024-06-18

Introduction

Berbagai negara di beberapa belahan dunia memiliki perbedaan kondisi perekonomiannya salah satunya dipengaruhi oleh ketersediaan pekerjaan. Di beberapa negara, masih banyak tingkat pengangguran yang membuat kondisi perekonomian di negara tersebut tidak stabil. Menurut sumber *Trading Economics*, di negara Afrika Selatan menyentuh hampir 33% untuk tingkat pengangguran, di negara India dengan 7,64%, di negara Kanada dengan 6,2% dan masih banyak lagi. Untuk itu, kami melakukan pencarian mendalam mengenai pekerjaan yang paling banyak dicari saat ini dalam membantu memperbaiki tingkat pengangguran di beberapa negara seperti tema SDGs ke 8 yaitu "*Decent Work and Economic Growth*". Belakangan tahun terakhir, pekerjaan di bidang Data Science semakin dibutuhkan oleh perusahaan-perusahaan besar karena dengan semakin berkembangnya teknologi dan bisnis maka diperlukan pengambilan keputusan yang lebih cerdas berdasarkan data yang tersedia. Kami menggunakan dataset "*Data Science Salary*" untuk menjelaskan tren gaji di bidang Data Science untuk tahun 2020 hingga 2024 dengan berfokus pada berbagai aspek ketenagakerjaan, termasuk pengalaman kerja, jabatan, dan lokasi perusahaan, data ini memberikan informasi tentang distribusi gaji dalam industri data science.

Permasalahan yang akan dibahas lewat analisa ini:

1) Tipe pekerjaan di bidang data science apa yang paling banyak dicari dan gajinya yang besar?
2) Negara mana yang memerlukan tenaga kerja paling banyak di bidang data science?
3) Bagaimana trend gaji pekerjaan data science dari tahun ke tahun (2020 hingga 2024)?
4) Bagaimana distribusi model kerja (remote, on-site, hybrid) bervariasi berdasarkan ukuran perusahaan (small, medium, large)?
5) Bagaimana distribusi jenis pekerjaan (full-time, part-time, freelancer, contract) di berbagai tingkat pengalaman (entry-level, mid-level, senior-level, executive-level)?

Kami menggunakan berbagai metode dan teknik analisis data untuk mengeksplorasi dan memahami pola serta tren dalam dataset. Beberapa metode utama yang kami terapkan adalah Exploratory Data Analysis (EDA) untuk memvisualisasikan distribusi dan hubungan antara variabel-variabel dalam dataset dalam bentuk bar plot serta menghitung frekuensi setiap kategori dalam variabel dan mengelompokkan data berdasarkan variabel tertentu dan menghitung rata-rata. Lalu kami juga memanfaatkan Plotly untuk membuat visualisasi

interaktif, seperti line chart, bar chart dan map yang memungkinkan eksplorasi data secara dinamis.

Analisis ini layak dibaca untuk para pembaca yang ingin mengetahui lebih dalam tentang bidang data science, untuk yang mau melamar kerja ataupun ingin memulai karir di bidang data science karena dapat menjadi wadah untuk wawasan serta informasi mengenai tren dan pola dalam industri data science. Bagi kami, yang merupakan mahasiswa data science, lewat analisis ini kami dapat memahami lebih lanjut mengenai faktor-faktor yang mempengaruhi gaji, model kerja, dan distribusi pekerjaan yang merupakan kunci untuk membuat keputusan karir kami kedepannya dengan lebih baik dan strategis.

Source: 1. Unemployment Rate menurut Trading Economics

2. Dataset Kaggle

Data Description

We have a total of 6600 records with 11 variables, namely job title, experience level, employment type, work models, work year, employee residence, salary, salary currency, salary in USD, company location, and company size.

1) work_year: Representing the specific year of salary data collection.
2) Experience_level: The level of work experience of the employees, categorized as EN (Entry-Level), EX (Experienced), MI (Mid-Level), SE (Senior).
3) Employment_type: The type of employment, labelled as FT (Full-Time), CT (Contractor), FL (Freelancer), PT (Part-Time).
4) Job_title: The job titles of the employees, such as “Applied Scientist”, “Data Quality Analyst”, etc.
5) Salary: The salary figures in their respective currency formats.
6) Salary_currency: The currency code representing the salary.
7) Salary_in_usd: The converted salary figures in USD for uniform comparison.
8) Company_location: The location of the companies, specified as country codes (e.g., “US” for the United States and “NG” for Nigeria).
9) Company_size: The size of the companies, classified as “L” (Large), “M” (Medium), and “S” (Small).
10) Work_models: Describes different working models, categorized as (Remote), (On-site), (Hybrid)
11) Employment_residence: The residence location of the employee.

Data Preprocessing

```
library(ggplot2)
library(dplyr)
library(outliers)
library(EnvStats)
library(plotly)
library(leaflet)
```

```r
data <- read.csv("data_science_salaries.csv")
```

Missing Value

```r
b = apply(data, 2, function(x) sum(is.na(x)))
b
```

```
##          job_title   experience_level    employment_type
work_models
##                  0                  0                  0
0
##          work_year employee_residence             salary
salary_currency
##                  0                  0                  0
0
##      salary_in_usd   company_location       company_size
##                  0                  0                  0
```

Disini, kita mencari tahu berapa jumlah missing value di semua kolom dari data yang kita pakai. Dari hasil output, bahwa **tidak ada missing value** di 11 kolom variabel.

Duplicated Data

```r
# Menyimpan ukuran awal data
size0 <- nrow(data)

# Menghapus duplikat
data <- data[!duplicated(data), ]

# Menyimpan ukuran setelah penghapusan duplikat
size1 <- nrow(data)

# Mencetak hasil
cat(paste(size0 - size1, "records out of", size0, "are duplicates"))
```

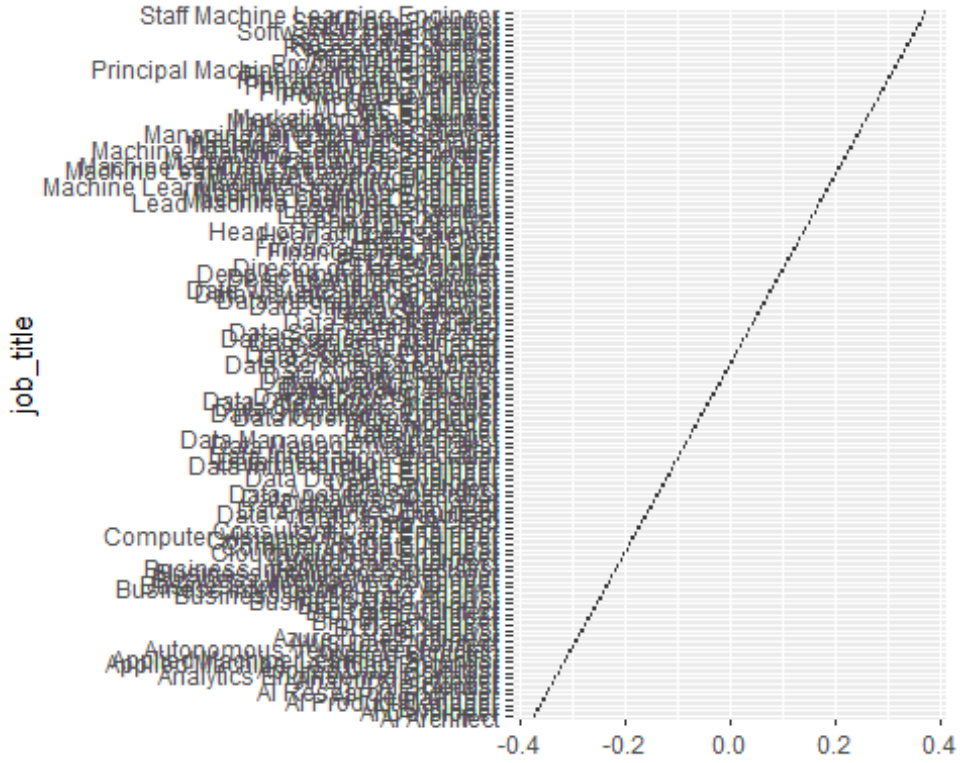
```
## 0 records out of 6599 are duplicates
```

Disini, kita tahu bahwa **tidak ada data yang terduplikasi** di 11 kolom variabel.

Outlier

1) job_title

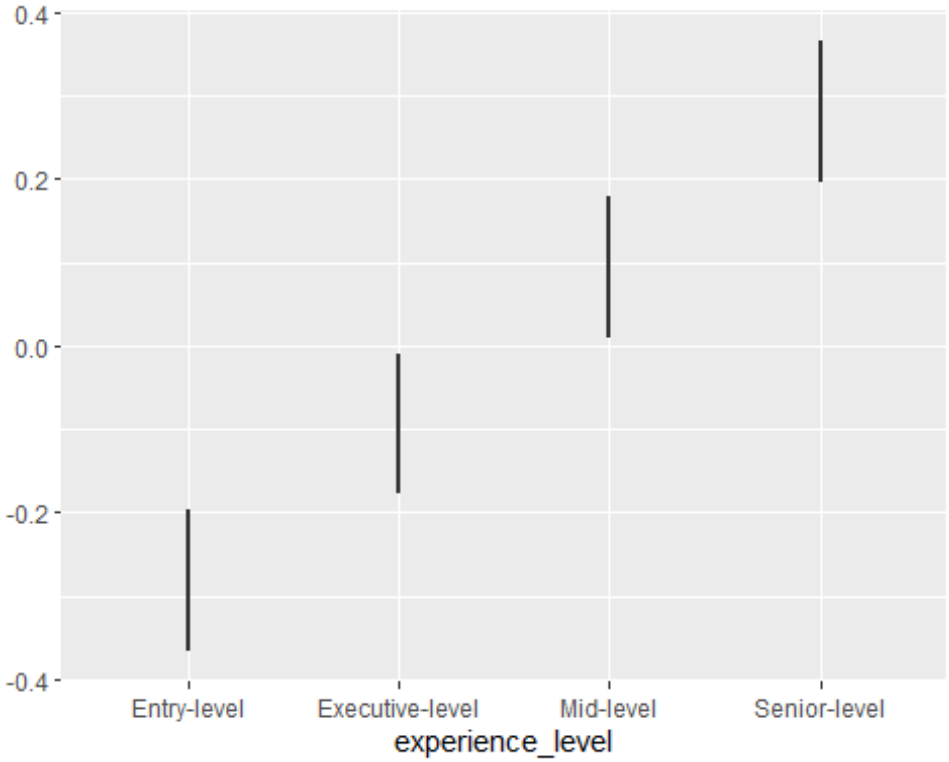
```r
ggplot(data, aes(y = job_title)) + geom_boxplot()
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel job_title, **tidak memiliki outliers** atau nilai pencilan karena tidak ada tanda bulat yang terletak jauh dari garis, data tersebar dengan rata. Dengan begitu variabel ini dapat digunakan dalam analisis lebih lanjut.

2) experience_level

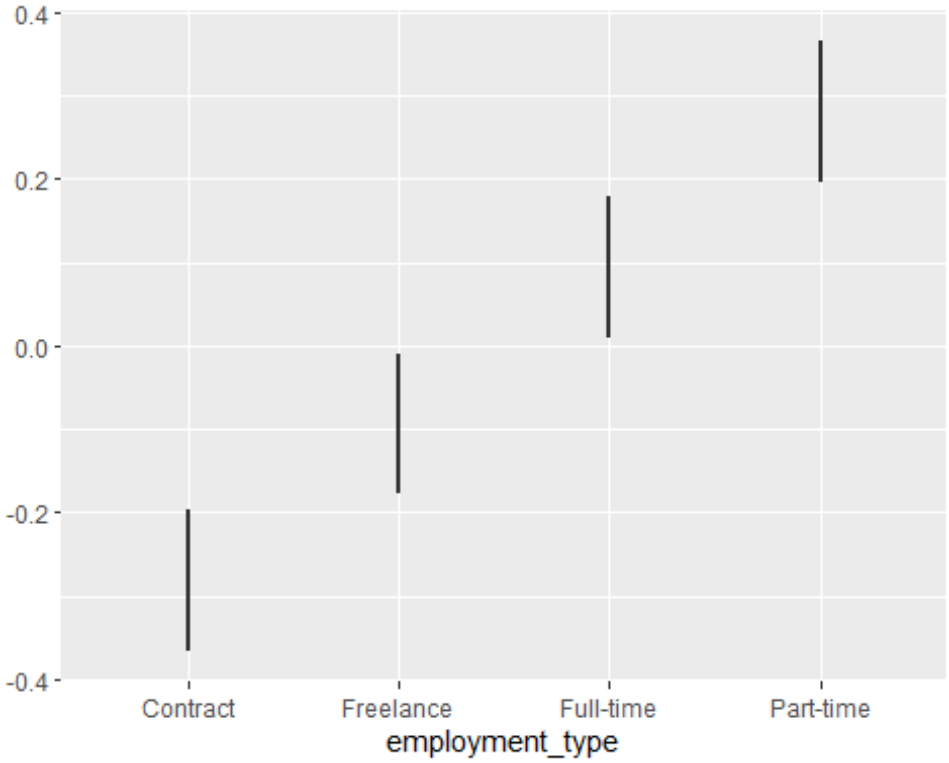
```
ggplot(data, aes(experience_level)) + geom_boxplot()
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel experience_level, **tidak memiliki outliers** atau nilai pencilan karena tidak ada tanda bulat yang terletak jauh dari garis, data tersebar dengan rata. Dengan begitu variabel dapat digunakan dalam analisis lebih lanjut.

3) employment_type

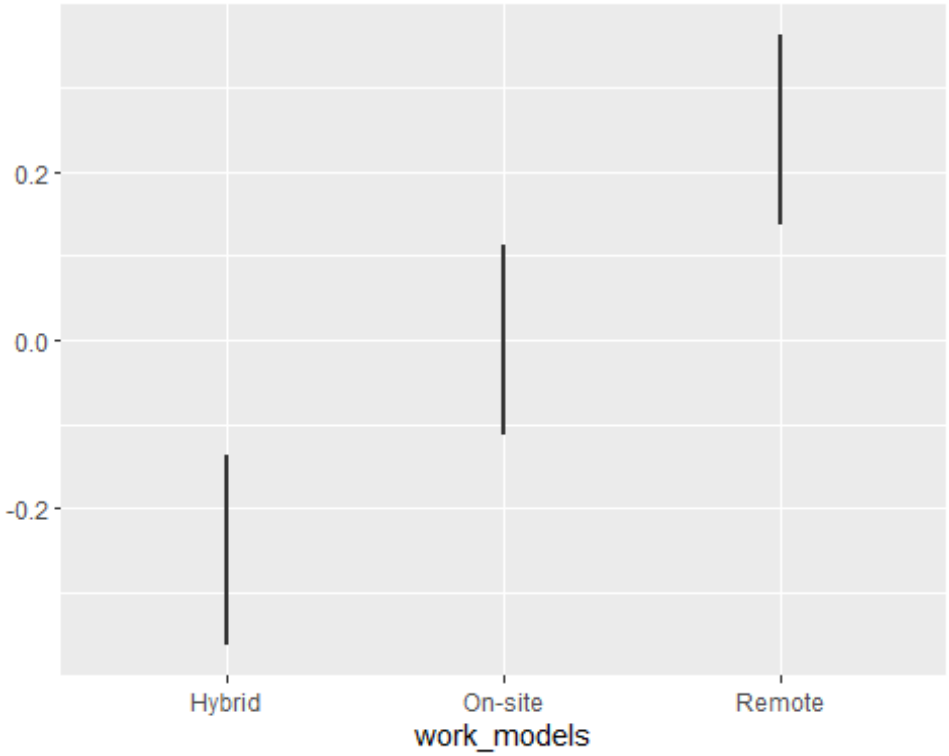
```
ggplot(data, aes(employment_type)) + geom_boxplot()
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel employment_type, **tidak memiliki outliers** atau nilai pencilan karena tidak ada tanda bulat yang terletak jauh dari garis, data tersebar dengan rata. Dengan begitu variabel dapat digunakan dalam analisis lebih lanjut.

4) work_models

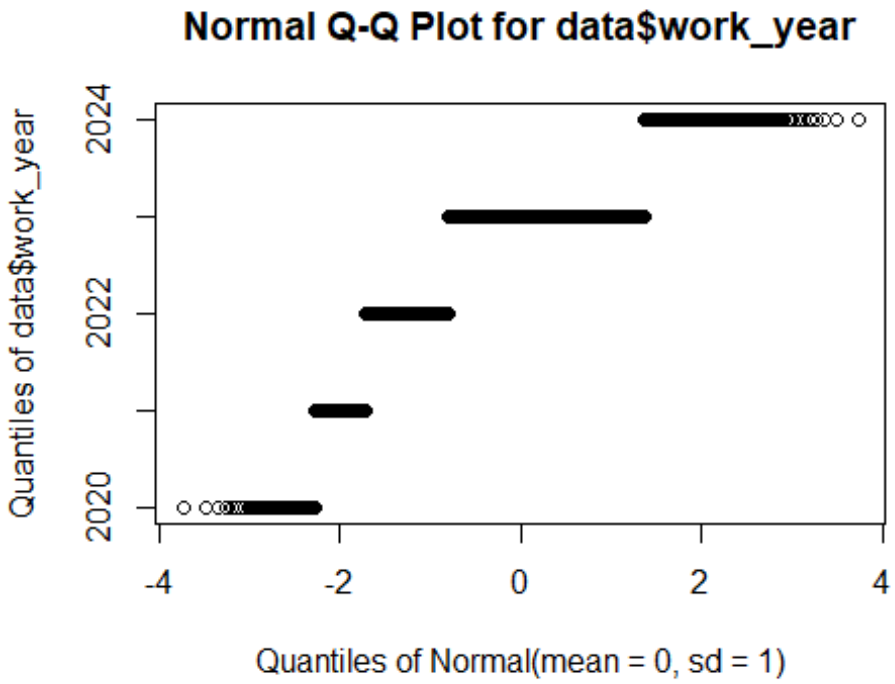
```
ggplot(data, aes(work_models)) + geom_boxplot()
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel work_models, **tidak memiliki outliers** atau nilai pencilan karena tidak ada tanda bulat yang terletak jauh dari garis, data tersebar dengan rata. Dengan begitu variabel dapat digunakan dalam analisis lebih lanjut.

5) work_year

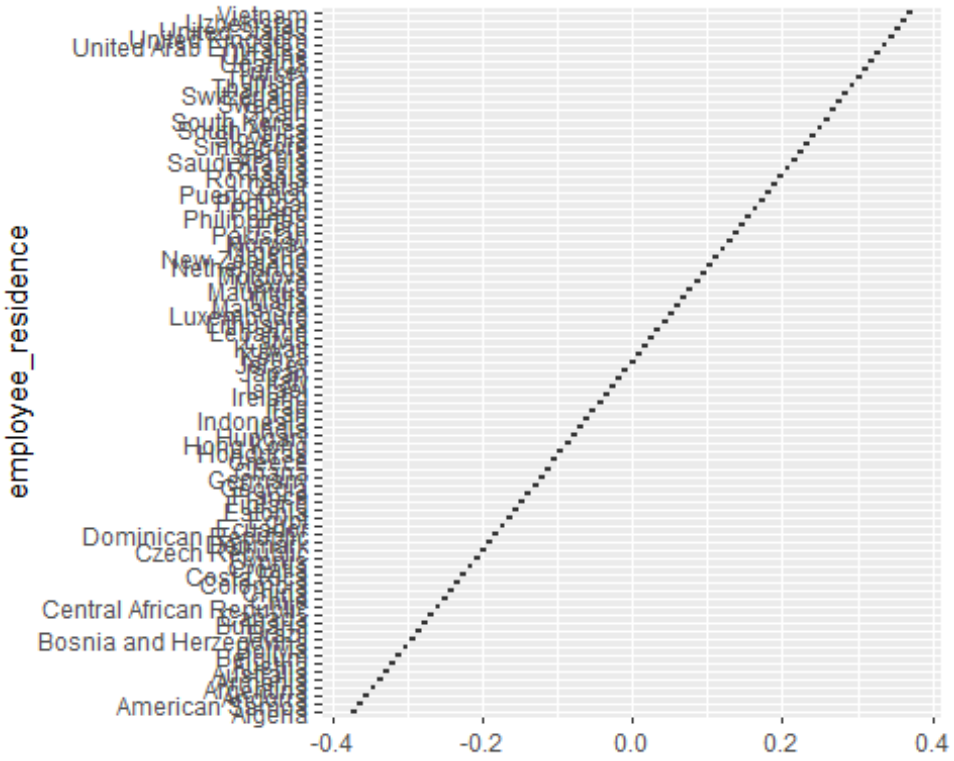
```
qqPlot(data$work_year)
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel work_year, **memiliki outliers** atau nilai pencilan karena ada tanda bulat yang terletak jauh dari garis grafiknya. Keberadaan outlier ini *perlu diperhatikan* dalam analisis lebih lanjut untuk memastikan bahwa hasilnya *tidak terpengaruh* secara signifikan oleh nilai nilai yang ekstrem.

6) employee_residence

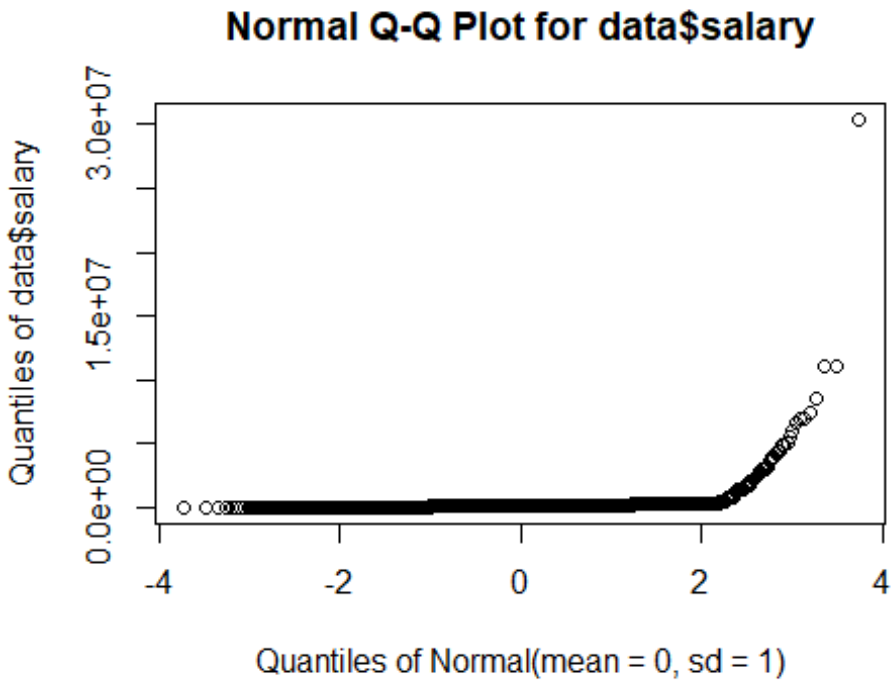
```
ggplot(data, aes(y = employee_residence)) + geom_boxplot()
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel employee_residence, **tidak memiliki outliers** atau nilai pencilan karena tidak ada tanda bulat yang terletak jauh dari garis, data tersebar dengan rata. Dengan begitu variabel ini dapat digunakan dalam analisis lebih lanjut.

7) salary

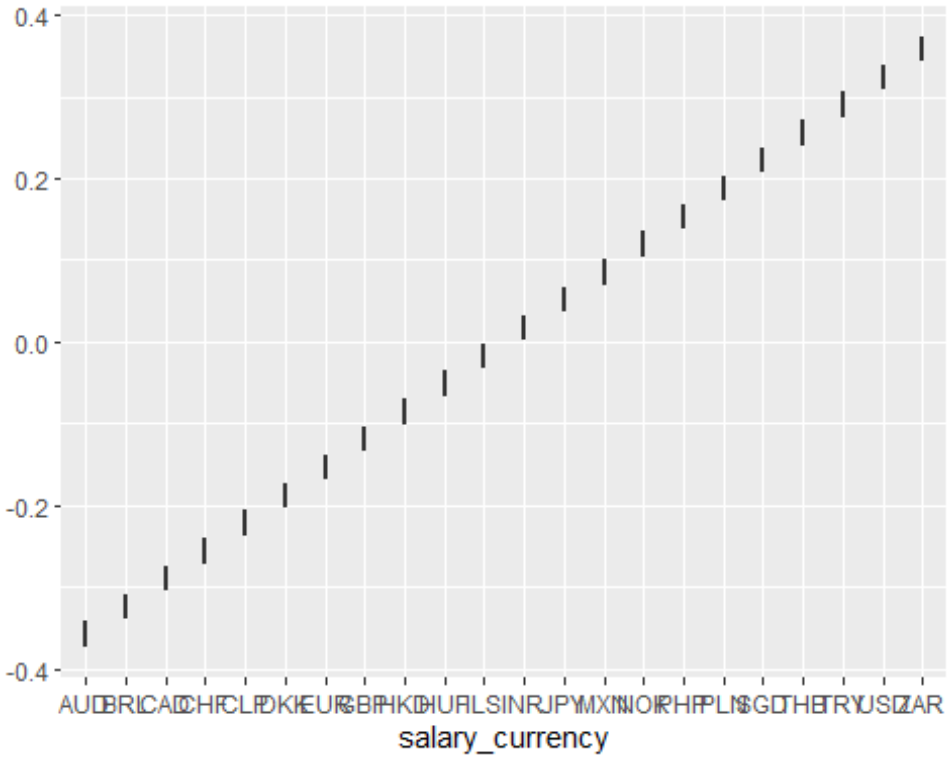
```
qqPlot(data$salary)
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel salary, **memiliki outliers** atau nilai pencilan karena ada tanda bulat yang terletak jauh dari garis grafiknya. Keberadaan outlier ini *perlu diperhatikan* dalam analisis lebih lanjut untuk memastikan bahwa hasilnya *tidak terpengaruh* secara signifikan oleh nilai nilai yang ekstrem.

8) salary_currency

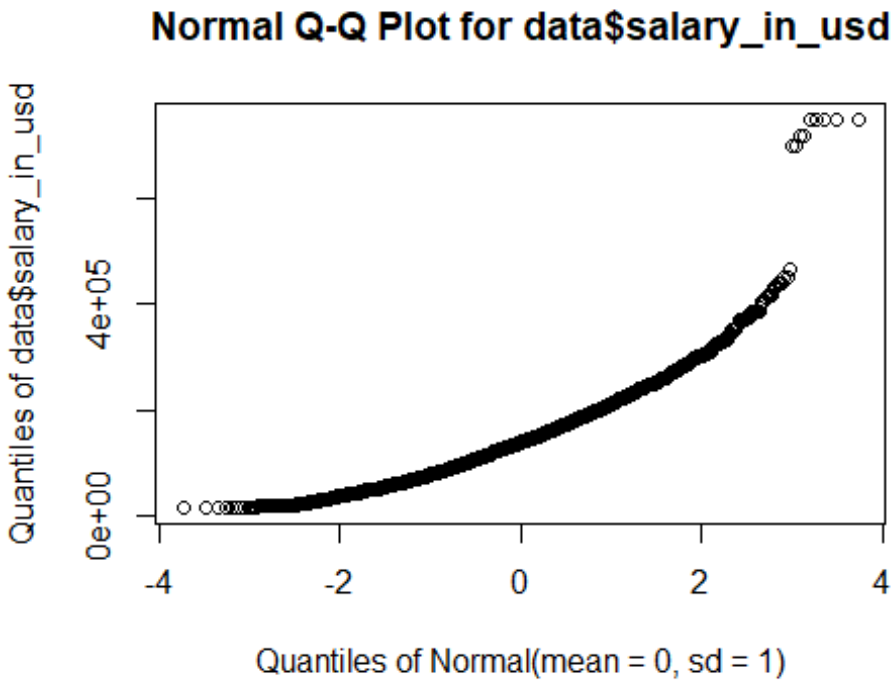
```
ggplot(data, aes(salary_currency)) + geom_boxplot()
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel salary_currency, **tidak memiliki outliers** atau nilai pencilan karena tidak ada tanda bulat yang terletak jauh dari garis, data tersebar dengan rata. Dengan begitu variabel ini dapat digunakan dalam analisis lebih lanjut.

9) salary_in_usd

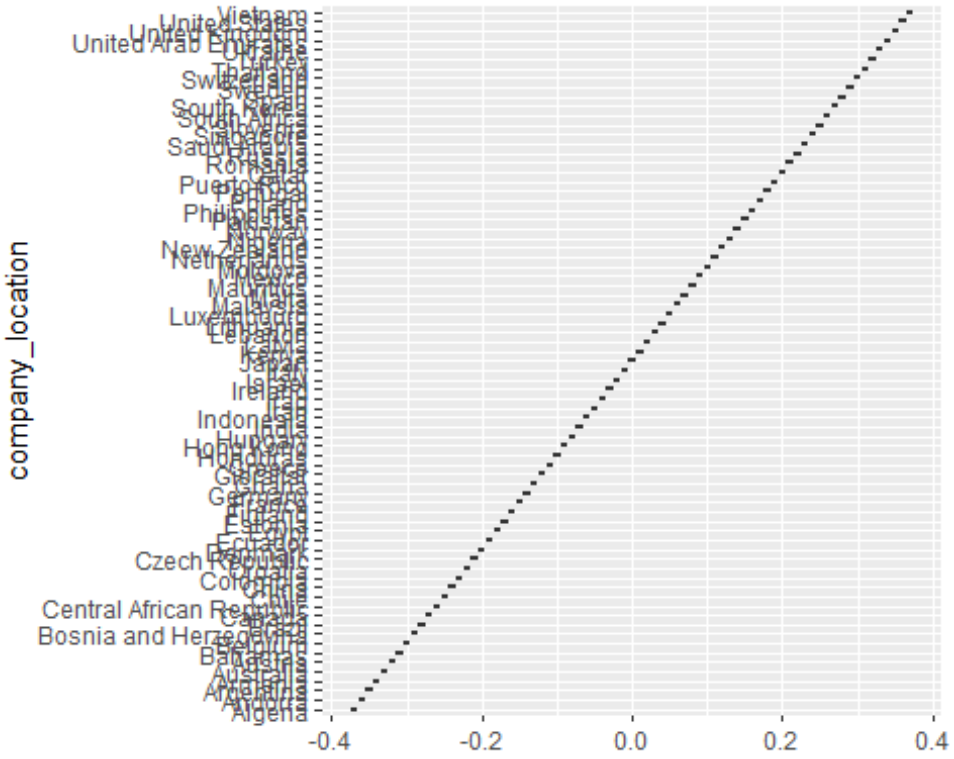
```
qqPlot(data$salary_in_usd)
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel salary_in_usd, **memiliki outliers** atau nilai pencilan karena ada tanda bulat yang terletak jauh dari garis grafiknya. Keberadaan outlier ini *perlu diperhatikan* dalam analisis lebih lanjut untuk memastikan bahwa hasilnya *tidak terpengaruh* secara signifikan oleh nilai nilai yang ekstrem.

10) company_location

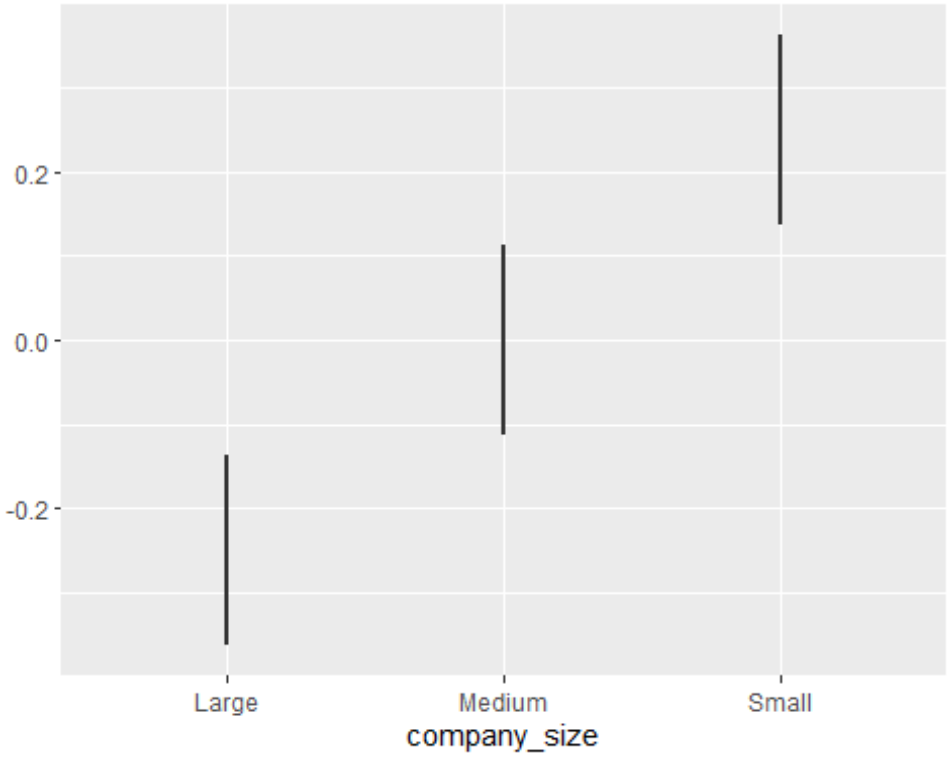
```
ggplot(data, aes(y = company_location)) + geom_boxplot()
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel company_location, **tidak memiliki outliers** atau nilai pencilan karena tidak ada tanda bulat yang terletak jauh dari garis, data tersebar dengan rata. Dengan begitu variabel ini dapat digunakan dalam analisis lebih lanjut.

11) company_size

```
ggplot(data, aes(company_size)) + geom_boxplot()
```



Dengan menggunakan function boxplot, bisa diketahui kalau di variabel company_size, **tidak memiliki outliers** atau nilai pencilan karena tidak ada tanda bulat yang terletak jauh dari garis, data tersebar dengan rata. Dengan begitu variabel ini dapat digunakan dalam analisis lebih lanjut.

KESIMPULAN OUTLIER KESELURUHAN

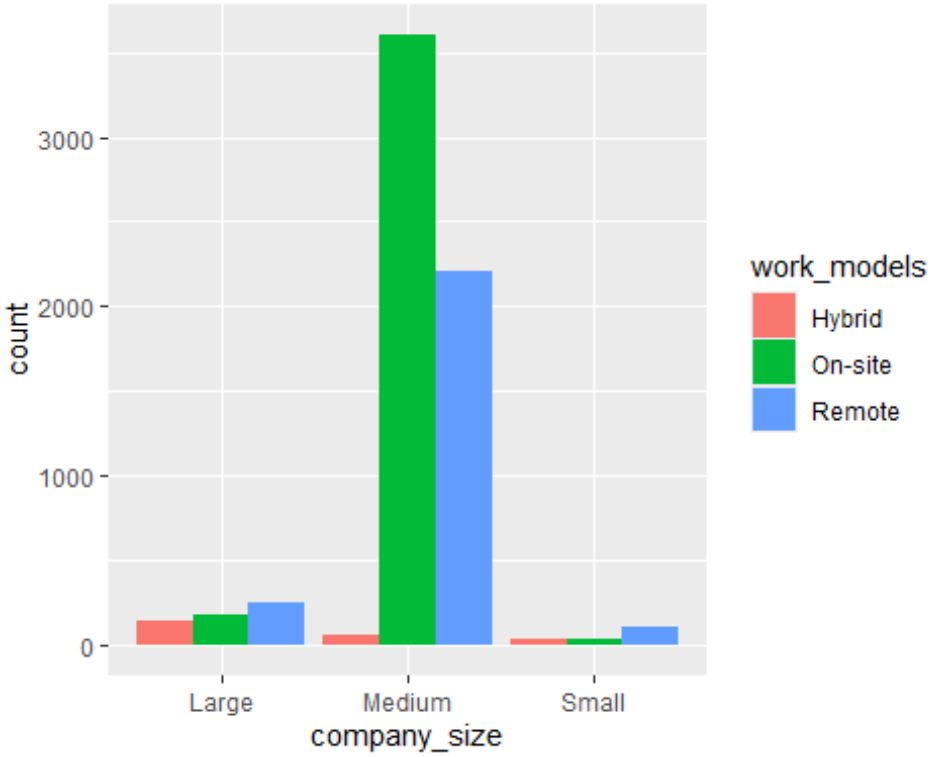
Berdasarkan analisis outlier yang telah dilakukan, dapat disimpulkan bahwa variabel yang *memiliki outliers* adalah **work_year**, **salary**, dan **salary_in_usd**. Variabel lainnya **(job_title, experience_level, employment_type, work_models, employee_residence, salary_currency, company_location, company_size)** *tidak memiliki* outliers yang signifikan. Hal ini menunjukkan bahwa sebagian besar data tersebar secara rata, namun beberapa variabel terkait dengan gaji dan tahun memiliki beberapa nilai yang menyimpang. Analisa outlier ini menggunakan **ggplot2** dan **qqPlot**, **ggplot2** untuk visualisasi yang kustomisasi dan konsisten, sedangkan **qqPlot** untuk mendeteksi outliers dan validasi distribusi data. Penggunaan kedua plot ini memungkinkan untuk perolehan infomasi yang *lebih mendalam dan akurat* dari dataset yang dianalisis.

Data Exploration

Examine Exploratory Visualization

1) work_models and company_size

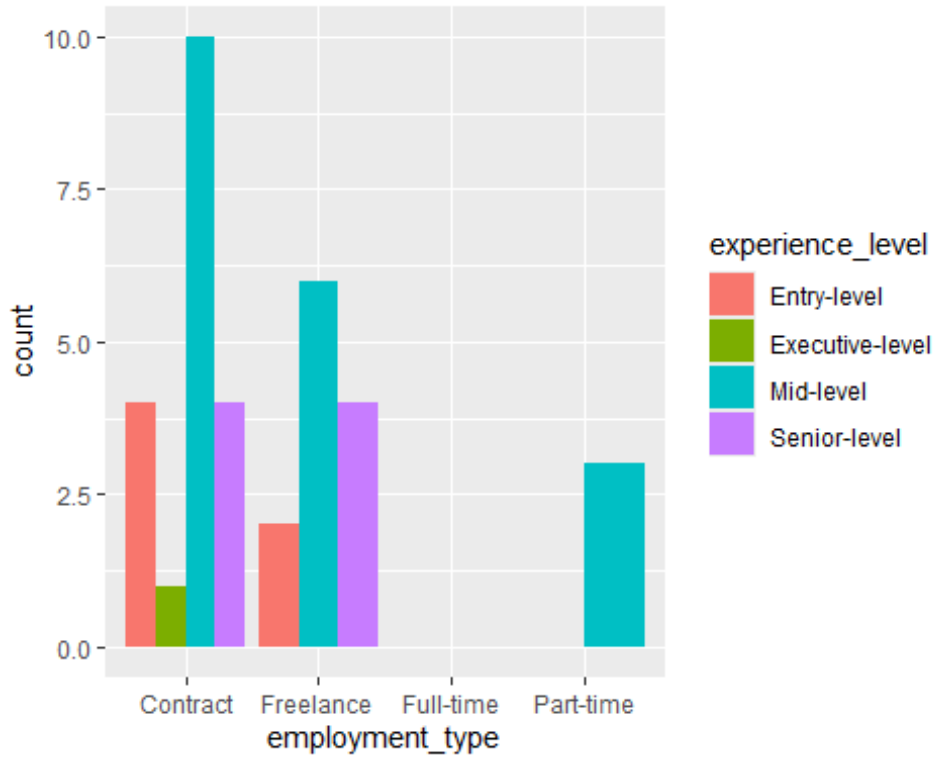
```
ggplot(data, aes(x = company_size, fill = work_models)) +
  geom_bar(position = "dodge")
```



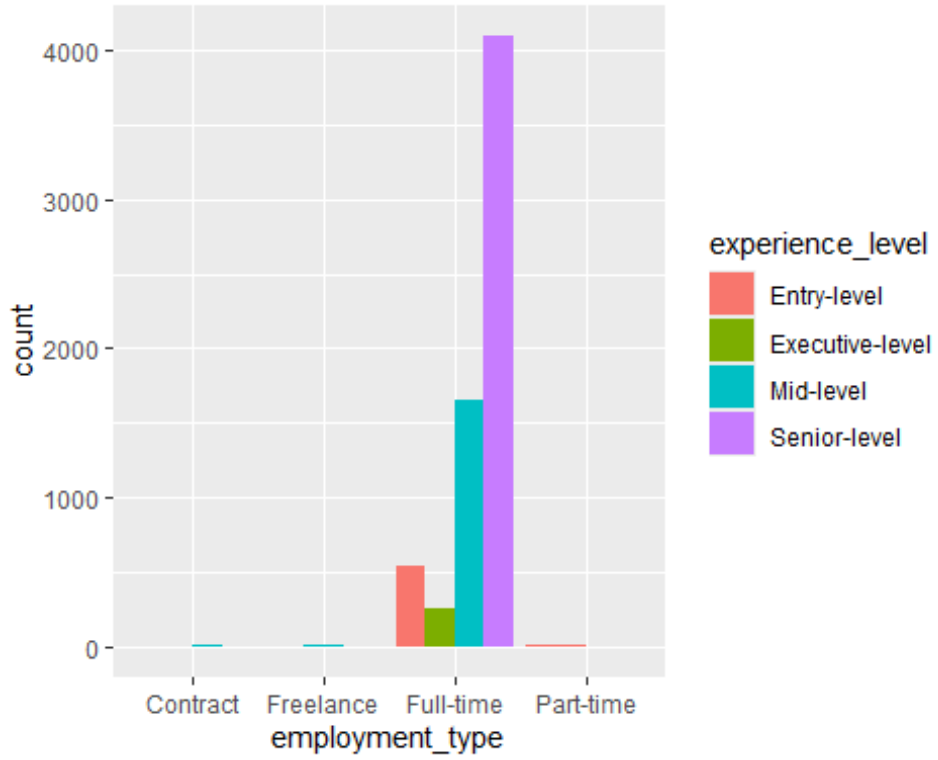
Dari grafik diatas, bisa dilihat bahwa perusahaan **dengan ukuran large, paling banyak pekerjanya bekerja secara remote** dan paling sedikit pekerjanya bekerja secara hybrid. Perusahaan dengan ukuran medium, paling banyak pekerjanya bekerja secara on-site dan paling sedikit pekerjanya bekerja secara hybrid. Perusahaan dengan ukuran small, paling banyak pekerjanya bekerja secara remote dan paling sedikit pekerjanya bekerja secara hybrid.

2) employment_type and experience_level

```
ggplot(data, aes(x = employment_type, fill = experience_level)) +
  geom_bar(position = "dodge") + scale_y_continuous(limits = c(0, 10))
```




```r
ggplot(data, aes(x = employment_type, fill = experience_level)) +
  geom_bar(position = "dodge")
```



Dari grafik diatas, bisa dilihat di jenis pekerjaan Contract, **pekerja di mid-level merupakan jumlah pekerja yang paling banyak dan pekerja di executive-level merupakan jumlah pekerja yang paling sedikit.** Di jenis pekerjaan Freelance, pekerja di mid-level merupakan jumlah pekerja yang paling banyak dan pekerja di entry-level merupakan jumlah pekerja yang paling sedikit, tidak ada pekerja executive-level di Freelance. Di jenis pekerjaan Part-time, pekerja di mid-level saja yang bekerja di pekerjaan Part-time ini. Di jenis pekerjaan Full-time, pekerja di senior-level merupakan jumlah pekerja yang paling banyak dan pekerja di executive-level merupakan jumlah pekerja yang paling sedikit.

Interactive Visualization

1) work_year and salary_in_usd

```
df <- data %>%
  group_by(work_year) %>%
  summarize(average_salary = mean(salary_in_usd))

plot_ly(
  data = df,
  x = ~work_year,
  y = ~average_salary,
  type = 'scatter',
  mode = 'lines'
)
```

Data poin yang kami visualisasikan adalah work_year dan salary_in_usd. Dengan menggunakan *line chart*, kita dapat mengetahui berapa banyak kenaikan dan penurunan gaji setiap tahun. Pada **2020-2021**, gaji yang dibayarkan kepada karyawan menurun sebesar 3.000 USD. Namun, **setelah 2021 hingga sekarang**, gaji telah meningkat drastis.

2) job-title and salary_in_usd

```
df <- data %>%
  group_by(job_title) %>%
  summarize(salary_in_usd = mean(salary_in_usd))


plot_ly(
  data = df,
  x = ~job_title,
  y = ~salary_in_usd,
  type = "bar"
)
```

Poin data yang kami visualisasikan adalah job_title dan salary_in_usd. Dengan menggunakan *bar chart*, kita dapat mengetahui berapa banyak yang diperoleh setiap pekerjaan dan pekerjaan mana yang memiliki gaji tertinggi. Pekerjaan dengan gaji tertinggi

adalah **Analytics Engineering Manager** dan pekerjaan dengan gaji terendah adalah **Data Analyst Lead**.

3) company_location and salary_in_usd

```
df <- data.frame(
  company_location = c("United States", "United Kingdom", "Canada",
"Portugal", "Poland", "Netherlands", "Italy", "Colombia", "Brazil", "Spain",
"Qatar", "Philippines", "Norway", "New Zealand", "Luxembourg", "Lithuania",
"Kenya", "South Korea", "Israel", "Hungary", "Hong Kong", "Croatia",
"Gibraltar", "Ghana", "Finland", "Ecuador", "Czech Repbulic", "Central
African Republic", "Bosnia and Herzegovina", "Armenia", "Andorra", "United
Arab Emirates", "Puerto Rico", "Pakistan", "Malaysia", "Iran", "Indonesia",
"Honduras", "Mauritius", "Bahamas", "Moldova", "India", "Egypt", "Vietnam",
"Turkey", "Romania", "Lebanon", "Ireland", "Germany", "Australia", "Ukraine",
"Thailand", "South Africa", "Singapore", "Saudi Arabia", "Slovenia",
"Sweden", "Russia", "Mexico", "Japan", "Nigeria", "Latvia", "Greece",
"France", "Estonia", "Denmark", "Argentina", "Switzerland", "Belgium",
"Malta", "Iraq", "Chile", "China", "Algeria", "Austria"),
  salary_in_usd = c(60000, 102898, 139833, 51513, 62139, 76969, 49997, 74775,
58569, 60888, 300000, 57670, 88642, 151634, 47609, 97611, 65000, 47000,
217332, 32140, 65058, 76726, 79976, 27000, 68519, 16000, 69479, 49216,
120000, 50000, 50745, 100000, 167500, 30000, 40000, 100000, 34208, 20000,
100000, 45555, 18000, 41699, 109367, 68000, 22314, 44713, 71750, 107634,
93536, 114673, 121333, 22971, 62935, 62783, 134999, 56186, 98791, 78028,
74865, 110822, 60444, 58626, 47411, 80905, 45993, 49403, 62000, 88770, 76865,
28369, 100000, 40038, 100000, 100000, 71355)
)

df_grouped <- df %>%
  group_by(company_location) %>%
  summarise(salary = mean(salary_in_usd, na.rm = TRUE))

fig <- plot_ly(
  data = df_grouped,
  type = 'choropleth',
  locations = ~company_location,
  locationmode = 'country names', # Menentukan bahwa lokasi adalah nama
negara
  z = ~salary,
  text = ~company_location,
  colorscale = "Blues"
)

fig
```

Peta data yang kami visualisasikan menunjukkan persebaran rata-rata gaji pekerja data science di berberapa negara, dengan **warna biru yang lebih gelap** menunjukkan gaji yang **lebih tinggi**. Beberapa negara seperti **Qatar**, **New Zealand**, dan **Israel**, yang ditandai dengan **warna biru gelap**, memiliki rata-rata gaji yang **sangat tinggi** misalnya, **Qatar**

memiliki rata-rata gaji ***300.000 USD***. Lalu negara-negara seperti **Kanada**, **Jerman**, dan **Australia** memiliki rata-rata gaji yang **cukup tinggi**, meskipun *tidak setinggi negara-negara dengan warna paling gelap*. Rata-rata gaji di negara-negara ini berkisar antara ***100.000*** hingga ***140.000 USD***. Dan terakhir, Negara-negara dengan warna yang **lebih terang**, seperti **Central African Republic**, **Ghana**, dan **Lebanon**, memiliki rata-rata gaji yang **lebih rendah**. Misalnya, **Central African Republic** memiliki rata-rata gaji sebesar ***16.000 USD****, menunjukkan perbedaan yang signifikan dalam salary dibandingkan dengan negara-negara dengan rata-rata gaji yang* lebih tinggi*.

Statitical Analysis

-> Analisis statistik seperti *chi-square test* dan juga *pearson correlation test* **tidak diperlukan** untuk menganalisis dataset yang kami gunakan. Hal ini dikarenakan **tujuan dari analisis** dataset kami adalah untuk **mendapatkan gambaran umum** mengenai karir di dunia data science, seperti bagaimana tipe kerjanya, bagaimana gajinya, sampai dengan korelasi antara jabatan dengan gaji. Semua ini bisa digambarkan dengan baik secara **visualisasi** saja, contohnya seperti menggunakan **ggplot2** sehingga para pembaca bisa mengerti mengenai gambaran umum dari karir data science yang mau disampaikan lewat analisis kami.

Discussion

Berdasarkan analisis yang dilakukan, kami mengetahui bahwa posisi **Analytics Engineering Manager** merupakan yang *paling dicari* dan memiliki *gaji tertinggi* dalam bidang data science diikuti oleh **Amerika Serikat** sebagai negara dengan *permintaan tinggi akan tenaga kerja data science*, terutama di perusahaan besar dan sedang. Tren gaji dalam beberapa tahun terakhir menunjukkan **peningkatan** dimana terlihat bahwa semakin banyak orang yang bergabung dalam bidang data science seiring berjalannya waktu. Puncaknya terlihat pada tahun **2023**, sementara tahun **2024** baru dimulai beberapa bulan, sehingga jumlah orang yang bergabung *tidak sebanyak* tahun **2023**. Distribusi work model juga menunjukkan bahwa **large company** cenderung mendukung lebih banyak *pekerja remote*, sementara **perusahaan sedang** lebih condong pada *pekerja onsite*. Pekerjaan **full-time** juga *mendominasi* dengan berbagai **experience level**, sedangkan pekerjaan **part-time** dan **contract lebih terbatas** dan sesuai dengan **experience level tertentu**.

Conclusion

Analisis ini memberikan wawasan yang mendalam tentang *tren gaji para profesional Data Science* pada tahun **2020-2024**, termasuk tren pekerjaan data science berdasarkan tingkat **pengalaman**, **jabatan**, dan **ukuran perusahaan**. Dengan memahami tren ini, pencari kerja dapat membuat keputusan karir yang lebih strategis kedepannya, sementara penyedia kerja dapat mengembangkan strategi kompensasi yang kompetitif untuk *menarik dan mempertahankan pekerja*. Data ini juga bermanfaat bagi pemerintah dalam merumuskan kebijakan tenaga kerja yang dapat **mendukung pertumbuhan sektor Data Science**. Selain itu, pemerintah dapat menggunakan analisis ini untuk mengidentifikasi kebutuhan pelatihan dan pendidikan yang relevan, guna mempersiapkan angkatan kerja yang siap menghadapi perkembangan teknologi di masa depan.