

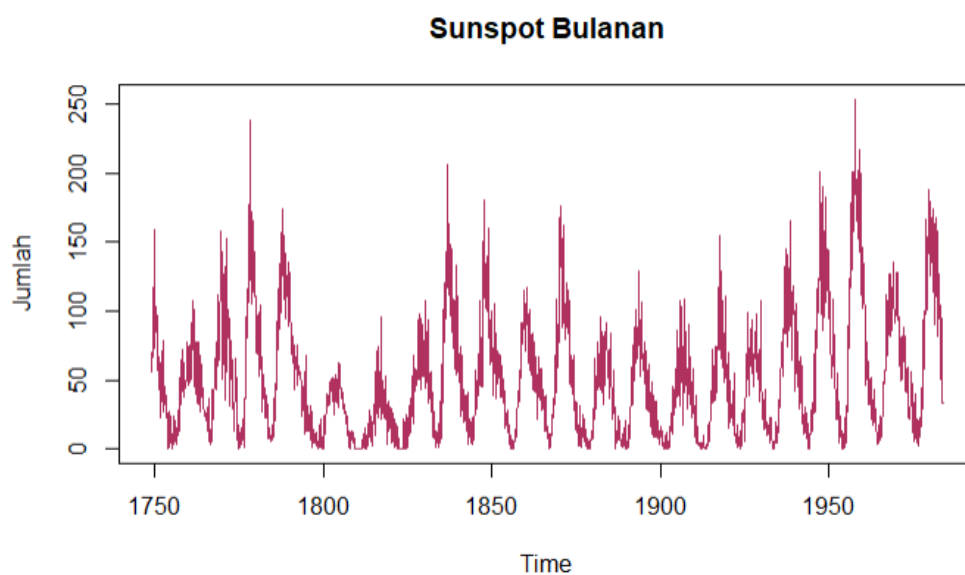
No 1.

## Pengolahan Data Sunspot Bulanan dengan Deret Waktu di R

Program ini bertujuan untuk memvisualisasikan dan memahami perilaku jumlah bintik matahari (sunspot) yang tercatat setiap bulan dari tahun 1749 hingga 1984. Data sunspot penting karena berkaitan dengan aktivitas matahari, yang dapat memengaruhi iklim bumi, komunikasi radio, dan sistem satelit.

# Sunspot Bulanan

```
1 ---
2 title: "sunspot.month"
3 author: "Reva Anwar"
4 date: "2025-04-20"
5 output: html_document
6 ---
7
8 ```{r}
9 url <- "C:\\Users\\HP\\Downloads\\monthly-sunspots.csv"
10 sunspot_month <- read.csv(url)
11
12
13 ```{r}
14 sun_ts <- ts(sunspot_month$Sunspots, start = c(1749, 1), frequency=12)
15 plot(sun_ts, main = "Sunspot Bulanan", ylab = "Jumlah", col="maroon")
16
```

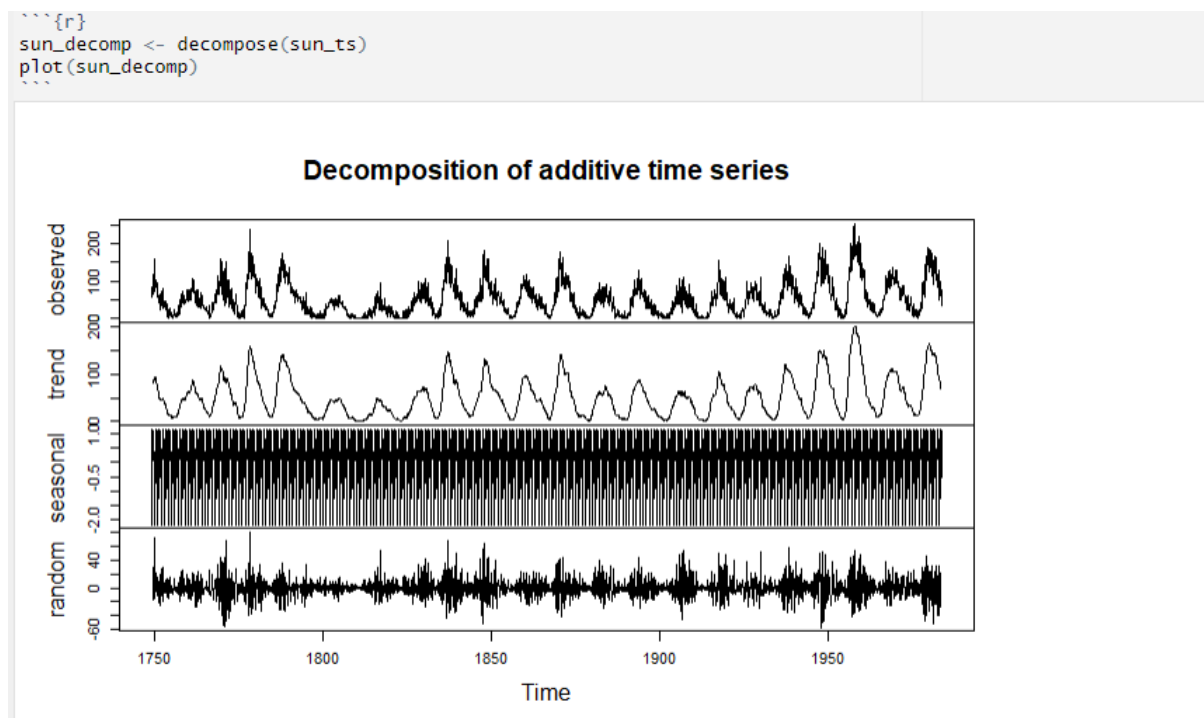


Grafik memperlihatkan pola naik-turun yang berulang secara berkala. Puncak (jumlah sunspot tertinggi) dan lembah (jumlah sunspot terendah) muncul secara teratur. Ini sesuai dengan

fenomena alami yang dikenal sebagai **siklus matahari**, yang memiliki rata-rata durasi sekitar **11 tahun** per siklus. Tinggi puncak sunspot tidak selalu sama dari satu siklus ke siklus berikutnya. Beberapa puncak memiliki nilai di atas 200 sunspot per bulan (terjadi pada sekitar tahun 1780-an dan 1950-an), sementara puncak lainnya relatif lebih rendah. Aktivitas matahari tidak selalu stabil. Terdapat periode ketika jumlah sunspot sangat rendah, misalnya antara tahun 1800–1820, yang berhubungan dengan fenomena yang dikenal sebagai **Dalton Minimum**, periode aktivitas matahari rendah.

- **Tahun 1750–1800:**  
Aktivitas sunspot cukup aktif dengan beberapa siklus yang kuat dan puncak yang tinggi.
- **Tahun 1800–1830:**  
Terjadi penurunan jumlah sunspot secara signifikan. Ini memperlihatkan periode minimum aktivitas yang cukup lama, diidentifikasi sebagai bagian dari **Dalton Minimum**.
- **Tahun 1830–1950:**  
Terjadi kembalinya aktivitas matahari dengan pola siklus yang lebih stabil dan kuat. Banyak puncak tinggi terjadi pada periode ini.
- **Tahun 1950–1980:**  
Aktivitas sunspot kembali meningkat drastis, dengan beberapa puncak aktivitas yang paling tinggi sepanjang seri ini. Ini menunjukkan fase matahari yang sangat aktif.

## Dekomposisi Time Series



Pada program yang dijalankan, dilakukan proses dekomposisi terhadap data deret waktu jumlah sunspot bulanan dari tahun 1749 hingga 1984. Program ini menggunakan fungsi `decompose()` dari R yang bertujuan untuk memisahkan data time series menjadi beberapa

komponen utamanya, yaitu **observed**, **trend**, **seasonal**, dan **random**. Hasil dari dekomposisi ini kemudian divisualisasikan ke dalam empat grafik terpisah untuk memudahkan analisis.

Komponen pertama yang ditampilkan adalah grafik **Observed**, yaitu data asli jumlah sunspot bulanan sepanjang periode pengamatan. Grafik ini memperlihatkan adanya pola fluktuasi yang berulang dengan variasi intensitas dari waktu ke waktu. Kita bisa melihat adanya siklus periodik di mana jumlah sunspot mengalami kenaikan dan penurunan dalam interval waktu yang relatif teratur, menandakan adanya pola berulang alami dalam aktivitas matahari.

Selanjutnya, grafik **Trend** menggambarkan pergerakan tren jangka panjang dari aktivitas sunspot. Terlihat bahwa tren ini tidak statis, melainkan mengalami perubahan yang cukup signifikan antar-dekade. Ada periode-periode di mana jumlah rata-rata sunspot meningkat, seperti pada pertengahan abad ke-20, serta periode di mana aktivitasnya menurun. Hal ini menunjukkan bahwa selain pola musiman, aktivitas matahari juga dipengaruhi oleh perubahan jangka panjang yang mungkin berkaitan dengan siklus yang lebih besar dalam dinamika internal matahari.

Pada bagian **Seasonal**, terlihat pola musiman yang sangat stabil dan berulang. Fluktuasi yang cepat dan tajam dari komponen ini menunjukkan adanya siklus musiman yang kuat. Siklus ini sesuai dengan siklus aktivitas matahari yang dikenal memiliki periode sekitar 11 tahun. Pola seasonal ini tampak hampir tidak berubah sepanjang rentang waktu analisis, menunjukkan bahwa meskipun terjadi perubahan tren jangka panjang, pola musiman tetap konsisten dan kuat.

Sementara itu, grafik terakhir, yaitu **Random**, memperlihatkan komponen sisa yang tidak dapat dijelaskan oleh tren maupun pola musiman. Komponen acak ini tampak lebih kecil dibandingkan dengan komponen lainnya, meskipun terdapat beberapa fluktuasi yang cukup tajam. Ini menunjukkan bahwa sebagian besar variasi pada data sunspot dapat dijelaskan oleh kombinasi tren jangka panjang dan pola musiman, sementara komponen acak hanya memberikan kontribusi yang relatif kecil terhadap keseluruhan variasi.

Secara keseluruhan, dekomposisi ini memperlihatkan bahwa data jumlah sunspot bulanan memiliki struktur yang sangat teratur, dengan pola musiman yang kuat dan konsisten, serta adanya fluktuasi tren jangka panjang yang cukup berarti. Komponen random yang tidak terlalu dominan menunjukkan bahwa data ini relatif stabil dan cocok untuk dilakukan pemodelan lebih lanjut, seperti peramalan menggunakan metode ARIMA musiman atau model berbasis trend-seasonal decomposition. Analisis ini memberikan pemahaman yang lebih dalam tentang perilaku aktivitas matahari dari waktu ke waktu dan membuka peluang untuk membuat prediksi yang lebih akurat terhadap fenomena alam yang dipengaruhi oleh siklus matahari.

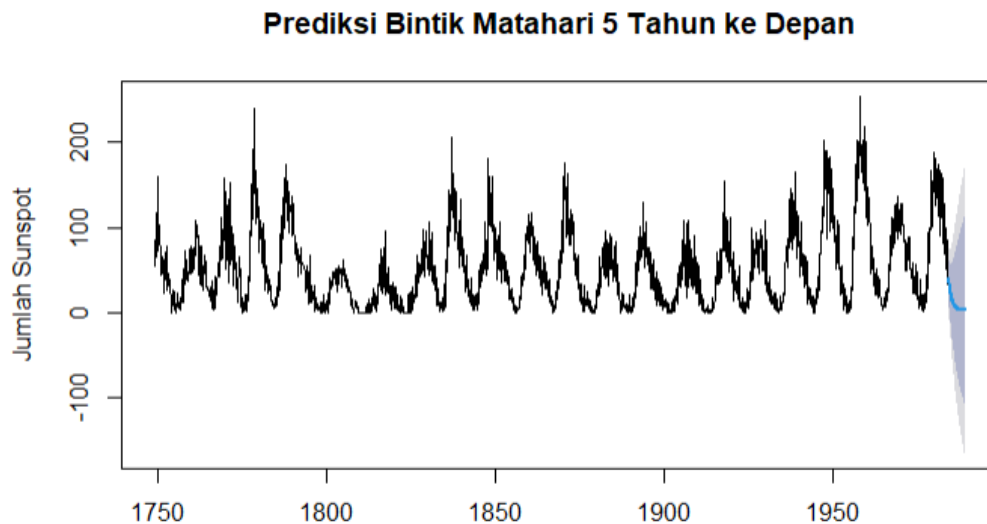
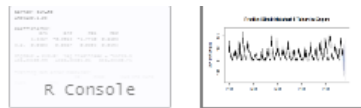
## Prediksi Bintik Matahari

Program ini memberikan alat analisis yang sangat berguna untuk memahami pola jumlah bintik matahari dan memperkirakan bagaimana polanya akan berlanjut di masa depan. Dengan model ARIMA yang dihasilkan, diharapkan prediksi bintik matahari menjadi lebih akurat, sehingga dapat digunakan untuk berbagai aplikasi ilmiah, khususnya dalam studi terkait aktivitas matahari dan dampaknya terhadap bumi.

```

##{r}
fit_arima <- auto.arima(sun_ts)
summary(fit_arima)
forecast_arima <- forecast(fit_arima, h = 60)
plot(forecast_arima, main = "Prediksi Bintik Matahari 5 Tahun ke Depan", ylab = "Jumlah Sunspot")

```

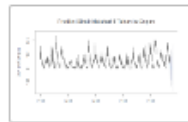
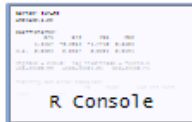


Grafik di atas menunjukkan hasil prediksi jumlah bintik matahari selama lima tahun ke depan menggunakan model ARIMA yang telah dibangun sebelumnya. Data historis jumlah bintik matahari ditampilkan dengan garis berwarna hitam, sedangkan prediksi untuk lima tahun berikutnya mulai terlihat di bagian akhir garis, dilengkapi dengan area berwarna biru yang merepresentasikan interval kepercayaan prediksi.

Berdasarkan pola historis, kita bisa mengamati adanya fluktuasi periodik atau siklus yang konsisten dalam aktivitas bintik matahari, dengan puncak-puncak dan lembah-lembah yang terjadi secara teratur sepanjang waktu. Ini mengindikasikan adanya karakteristik musiman yang kuat dalam data. Pada bagian prediksi, model ARIMA memperkirakan bahwa jumlah bintik matahari akan cenderung menurun dalam beberapa tahun pertama ke depan, ditandai dengan arah garis yang sedikit menurun, sebelum kemudian stabil.

Area berwarna biru memperlihatkan ketidakpastian prediksi. Semakin jauh ke masa depan, semakin lebar rentang kepercayaannya, yang mencerminkan meningkatnya ketidakpastian dalam prediksi. Hal ini normal dalam peramalan deret waktu, karena prediksi jangka panjang lebih rentan terhadap kesalahan prediksi dibandingkan prediksi jangka pendek.

Meski begitu, secara keseluruhan, model ini menunjukkan kecenderungan bahwa siklus bintik matahari tetap berlanjut dengan fluktuasi periodik, meskipun dalam periode prediksi terjadi penurunan jumlah secara umum dibandingkan beberapa puncak besar yang terlihat di masa lalu. Ini penting untuk berbagai bidang, mulai dari astronomi hingga cuaca antariksa, mengingat aktivitas matahari berpengaruh terhadap iklim bumi dan teknologi berbasis satelit.



```
Series: sun_ts
ARIMA(2,1,2)

Coefficients:
      ar1      ar2      ma1      ma2
    1.3467 -0.3963 -1.7710  0.8103
s.e.  0.0303  0.0287  0.0205  0.0194

sigma^2 = 243.8: log likelihood = -11745.5
AIC=23500.99 AICc=23501.01 BIC=23530.71

Training set error measures:
              ME      RMSE      MAE MPE MAPE      MASE      ACF1
Training set -0.02672716 15.60055 11.02575 NaN  Inf  0.4775401 -0.01055012
```

Berdasarkan hasil analisis data deret waktu sun\_ts, model ARIMA(2,1,2) telah diterapkan untuk memodelkan dan memprediksi jumlah bintik matahari. Model ini menunjukkan bahwa data awal tidak stasioner, sehingga perlu dilakukan differencing satu kali agar pola fluktuasi data lebih stabil dan dapat dianalisis secara statistik dengan baik. Setelah proses ini dilakukan, model ARIMA yang terbentuk menggunakan dua parameter untuk komponen autoregressive (AR) dan dua parameter untuk komponen moving average (MA), mencerminkan bahwa nilai saat ini sangat dipengaruhi oleh dua nilai sebelumnya dan dua kesalahan sebelumnya dalam data.

Koefisien AR pertama bernilai positif sebesar 1.3467, menunjukkan bahwa ada pengaruh positif dan cukup kuat dari nilai lag pertama terhadap nilai saat ini. Sebaliknya, koefisien AR kedua bernilai negatif sebesar -0.3963, yang mengindikasikan adanya efek pembalik atau penyesuaian dari dua waktu sebelumnya. Sementara itu, pada bagian moving average, koefisien MA pertama sebesar -1.7710 menunjukkan bahwa kesalahan pada satu waktu sebelumnya memberikan dampak negatif yang kuat pada nilai sekarang, dan MA kedua yang bernilai 0.8103 mengindikasikan bahwa kesalahan dari dua periode sebelumnya memberikan pengaruh positif. Seluruh koefisien ini disertai dengan nilai standar error yang cukup kecil, menandakan bahwa parameter-parameter tersebut relatif signifikan secara statistik.

Nilai log likelihood dari model ini adalah -11745.5, dan ukuran-ukuran statistik seperti AIC sebesar 23500.99, AICc sebesar 23501.01, dan BIC sebesar 23530.71 digunakan untuk menilai seberapa baik model ini cocok dengan data. Meskipun angka-angka ini hanya bermakna saat dibandingkan dengan model lain, tidak ada indikasi bahwa model ini buruk secara umum.

Ketika melihat hasil evaluasi terhadap data pelatihan, model ini menunjukkan nilai kesalahan rata-rata (ME) yang sangat kecil, yaitu sekitar -0.0267, yang artinya secara umum prediksi model tidak terlalu meleset ke arah manapun secara konsisten. Nilai RMSE (Root Mean Squared Error) sebesar 15.60055 dan MAE (Mean Absolute Error) sebesar 11.02575 memberikan gambaran tentang seberapa besar rata-rata kesalahan prediksi dalam satuan aslinya, yang dalam konteks ini masih dianggap cukup masuk akal. Namun, nilai MPE dan MAPE tidak tersedia (ditunjukkan sebagai NaN dan Inf), kemungkinan besar karena terdapat

nilai aktual yang nol dalam data, yang membuat pembagi dalam rumus metrik tersebut menjadi tidak valid. Meskipun demikian, nilai MASE sebesar 0.4775 justru memberikan informasi positif, karena nilai tersebut lebih kecil dari 1, menandakan bahwa model ARIMA ini lebih baik daripada model naive yang hanya menggunakan nilai sebelumnya sebagai prediksi. ACF1 atau autokorelasi residual lag ke-1 bernilai sangat kecil, yaitu -0.0105, yang menunjukkan bahwa sisa kesalahan dari model ini sudah bersifat acak dan tidak menunjukkan pola tertentu yang tertinggal.

Visualisasi hasil prediksi untuk lima tahun ke depan menunjukkan kelanjutan dari pola siklus bintik matahari dengan rentang ketidakpastian yang semakin lebar di masa depan. Ini adalah hal yang wajar dalam prediksi deret waktu, karena semakin jauh ke depan, semakin tinggi tingkat ketidakpastiannya. Grafik menunjukkan prediksi nilai tetap mengikuti pola naik turun yang mirip dengan tahun-tahun sebelumnya, namun disertai dengan bayangan abu-abu di sekitar garis prediksi, yang merupakan interval kepercayaan. Area ini semakin melebar seiring waktu, menandakan bahwa prediksi model menjadi semakin tidak pasti seiring menjauhnya dari data historis.

Secara keseluruhan, model ARIMA(2,1,2) yang diterapkan pada data `sun_ts` dapat dianggap cukup baik dalam menggambarkan pola historis dan memberikan prediksi yang wajar untuk lima tahun ke depan. Meskipun ada keterbatasan pada beberapa metrik evaluasi karena nilai nol dalam data, model ini tetap memberikan hasil yang konsisten dan layak untuk digunakan dalam keperluan peramalan bintik matahari jangka pendek.

## Output

```
> url <- "C:\\Users\\HP\\Downloads\\monthly-sunspots.csv"
> sunspot_month <- read.csv(url)
> sun_ts <- ts(sunspot_month$sunspots, start = c(1749, 1), frequency=12)
> plot(sun_ts, main = "Sunspot Bulanan", ylab = "Jumlah", col="maroon")
> sun_decomp <- decompose(sun_ts)
> plot(sun_decomp)
> fit_arima <- auto.arima(sun_ts)
> summary(fit_arima)
Series: sun_ts
ARIMA(2,1,2)

Coefficients:
          ar1          ar2          ma1          ma2
      1.3467   -0.3963   -1.7710    0.8103
s.e.  0.0303    0.0287    0.0205    0.0194

sigma^2 = 243.8:  log likelihood = -11745.5
AIC=23500.99   AICc=23501.01   BIC=23530.71

Training set error measures:
              ME          RMSE          MAE  MPE  MAPE          MASE          ACF1
Training set -0.02672716 15.60055 11.02575  NaN   Inf  0.4775401 -0.01055012
> forecast_arima <- forecast(fit_arima, h = 60)
> plot(forecast_arima, main = "Prediksi Bintik Matahari 5 Tahun ke Depan", ylab = "Jumlah sunspot")
> |
```

Kesimpulan akhir dari hasil analisis ini menunjukkan bahwa model ARIMA(2,1,2) mampu memodelkan pola jumlah bintik matahari dengan cukup baik. Nilai kesalahan prediksi relatif rendah, dan sisa (residual) dari model tidak menunjukkan pola tertentu, yang menandakan model ini sudah menangkap struktur data secara efektif. Prediksi lima tahun ke depan menunjukkan kelanjutan pola musiman yang konsisten dengan data historis, meskipun

dengan ketidakpastian yang meningkat seiring waktu. Secara keseluruhan, model ini dapat diandalkan untuk melakukan peramalan jangka pendek terhadap jumlah bintik matahari.


No 2.

## Visualisasi Persebaran Pelanggan Olist dengan RStudio

Kumpulan Data Publik E-Commerce Brasil oleh Olist

# Import Dataset

```
1- ---
2- title: "Analisis Data Pelanggan Olist"
3- author: "Reva Anwar"
4- date: "2023-04-21"
5- output: html_document
6- ---
7-
8- ```{r}
9- data <- read_csv("C:\\Users\\HP\\Downloads\\olist_customers_dataset.csv\\olist_customers_dataset.csv")
10- head(data)
11- ```
```



customer_id	customer_unique_id	customer_zip_code_prefix
06b8999e2fba1a1fbc88172c00ba8bc7	861eff4711a542e4b93843c6dd7febb0	14409
18955e83d337fd6b2def6b18a428ac77	290c77bc529b7ac935b93aa66c333dc3	09790
4e7b3e00288586ebd08712fdd0374a03	060e732b5b29e8181a18229c7b0b2b5e	01151
b2b6027bc5c5109e529d4dc6358b12c3	259dac757896d24d7702b9acbbff3f3c	08775
4f2d8ab171c80ec8364f7c12e35b23ad	345ecd01c38d18a9036ed96c73b8d066	13056
879864dab9bc3047522c92c82e1212b8	4c93744516667ad3b8f1fb645a3116a4	89254

Jika melihat lima baris pertama yang ditampilkan, dapat dilihat bahwa struktur data cukup konsisten. Setiap pelanggan memiliki customer\_id, customer\_unique\_id, dan customer\_zip\_code\_prefix yang valid tanpa adanya nilai kosong ataupun data yang tidak terbaca. Ini menandakan bahwa file CSV telah berhasil diimpor dengan format yang sesuai ke dalam RStudio.

Melalui pengamatan awal terhadap hasil ini, dapat disimpulkan bahwa:

- Dataset pelanggan Olist memiliki struktur yang terorganisir dengan baik, memisahkan antara transaksi individual dan identitas pelanggan yang permanen.

- Informasi lokasi sudah tersedia walaupun dalam bentuk kode pos prefix, sehingga masih memungkinkan untuk analisis distribusi geografis secara agregat.
- Tidak ditemukan adanya anomali atau ketidaksesuaian data pada tampilan awal, yang berarti tahap loading data berhasil dan dataset siap untuk diproses ke tahap analisis berikutnya.

## Ringkasan Data

```
[r]
dim(data)
glimpse(data)
colSums(is.na(data))

[1] 99441    5
Rows: 99,441
Columns: 5
$ customer_id      <chr> "06b8999e2fba1a1fbc88172c00ba8bc7", "18955e83d337fd6b2def6b18a428ac77", "4e7b3...
$ customer_unique_id <chr> "861eff4711a542e4b93843c6dd7febb0", "290c77bc529b7ac933b93aa66c333dc3", "060e7...
$ customer_zip_code_prefix <chr> "14409", "09790", "01151", "08773", "13056", "89254", "04534", "35182", "81560...
$ customer_city      <chr> "Franca", "sao bernardo do campo", "sao paulo", "mogi das cruzes", "campinas"...
$ customer_state      <chr> "SP", "SP", "SP", "SP", "SC", "SP", "MG", "PR", "MG", "MG", "RJ", "SP", ...
```

Berdasarkan hasil yang ditunjukkan dari program analisis awal pada data pelanggan Olist, dapat disimpulkan bahwa dataset ini berisi sebanyak **99.441 baris data** dengan **5 atribut** informasi untuk setiap pelanggan. Ini menunjukkan bahwa data mencakup hampir seratus ribu entri pelanggan, yang merupakan jumlah cukup besar untuk memberikan gambaran menyeluruh tentang persebaran pelanggan Olist di seluruh wilayah operasionalnya. Dengan ukuran data sebesar ini, analisis yang dilakukan dapat dipercaya mewakili kondisi nyata pelanggan Olist.

Struktur data yang ditampilkan memperlihatkan bahwa kelima atribut yaitu `customer_id`, `customer_unique_id`, `customer_zip_code_prefix`, `customer_city`, dan `customer_state` semuanya bertipe karakter. Ini menunjukkan bahwa baik identitas pelanggan, informasi lokasi, maupun kode pos diperlakukan sebagai teks. Hal ini penting karena di dalam dunia nyata, kode pos seringkali mengandung angka nol di depan, yang bisa hilang jika diolah sebagai angka, sehingga penyimpanan dalam bentuk teks menjaga keakuratan informasi geografis pelanggan.

Dari contoh isi data yang diperlihatkan, terlihat bahwa pelanggan Olist tersebar di berbagai kota besar Brasil seperti *Franca*, *São Bernardo do Campo*, *São Paulo*, *Mogi das Cruzes*, dan *Campinas*. Kota-kota ini termasuk dalam negara bagian São Paulo (SP), yang memang dikenal sebagai pusat ekonomi utama di Brasil. Selain SP, beberapa pelanggan juga berasal dari negara bagian Minas Gerais (MG) dan Rio de Janeiro (RJ), yang menunjukkan bahwa cakupan geografis Olist cukup luas dan berfokus pada wilayah-wilayah dengan tingkat



aktivitas ekonomi tinggi. Ini menandakan bahwa Olist menargetkan pasar yang potensial dari sisi daya beli maupun jumlah penduduk.

Selain itu, hasil pemeriksaan missing value menunjukkan bahwa tidak terdapat nilai kosong (NA) pada seluruh kolom data. Ini berarti data pelanggan Olist memiliki **kualitas yang sangat baik** dalam hal kelengkapan informasi. Ketiadaan missing value ini mempercepat proses analisis lanjutan, karena tidak perlu dilakukan pembersihan data tambahan seperti pengisian atau penghapusan nilai kosong.

Secara umum, hasil dari program ini menggambarkan bahwa data pelanggan Olist:

- Sangat lengkap dan konsisten tanpa kehilangan informasi.
- Tersebar di banyak kota besar di Brasil, mendominasi pada negara bagian yang secara ekonomi strategis.
- Siap digunakan untuk berbagai macam analisis lanjutan seperti segmentasi pelanggan, analisis perilaku berdasarkan lokasi, optimasi distribusi logistik, atau prediksi tren pembelian.

Kondisi ini menguntungkan dari sisi bisnis, karena perusahaan dapat mengambil langkah strategis berdasarkan sebaran pelanggan aktual. Misalnya, mereka bisa memperkuat logistik di wilayah dengan konsentrasi pelanggan tinggi atau menjalankan kampanye pemasaran yang lebih spesifik untuk masing-masing negara bagian.

Melalui hasil program ini, dapat dikatakan bahwa tahap eksplorasi awal telah memberikan landasan kuat untuk memahami profil pelanggan Olist, serta membuktikan bahwa dataset ini dapat digunakan untuk mendukung berbagai keperluan analitik dan pengambilan keputusan bisnis.

## Kota dengan Jumlah Pelanggan Terbanyak (Top 10)

```
```{r}
data %>%
  count(customer_city, sort = TRUE) %>%
  slice_max(n, n = 10) %>%
  ggplot(aes(x = reorder(customer_city, n), y = n)) +
  geom_col(fill = "tomato") +
  coord_flip() +
  labs(title = "10 Kota dengan Jumlah Pelanggan Terbanyak", x = "kota", y = "Jumlah") +
  theme_minimal()
```
```



Dari grafik yang dihasilkan, terlihat sangat jelas bahwa **kota São Paulo** mendominasi jumlah pelanggan Olist dengan selisih yang sangat jauh dibandingkan kota-kota lainnya. São Paulo memiliki lebih dari 15.000 pelanggan, jauh meninggalkan kota berikutnya, yaitu **Rio de Janeiro**, yang jumlah pelanggannya mendekati angka 10.000. Hal ini menunjukkan bahwa basis pelanggan terbesar Olist terkonsentrasi di kota metropolitan São Paulo, yang memang merupakan pusat ekonomi terbesar di Brasil.

Setelah São Paulo dan Rio de Janeiro, di posisi ketiga terdapat **Belo Horizonte**, disusul oleh **Brasília** dan **Curitiba**. Meskipun jumlah pelanggan di kota-kota ini jauh lebih sedikit dibandingkan dua kota teratas, tetap saja mereka termasuk dalam kelompok kota penting dalam basis pelanggan Olist.

Kota-kota berikutnya dalam daftar adalah **Campinas, Porto Alegre, Salvador, Guarulhos**, dan **São Bernardo do Campo**. Kota-kota ini memiliki jumlah pelanggan yang cukup berdekatan satu sama lain, meskipun masih jauh lebih rendah dibandingkan São Paulo dan Rio de Janeiro.

Distribusi pelanggan yang ditunjukkan dalam grafik ini menggambarkan pola yang sangat khas, yaitu bahwa sebagian besar pelanggan Olist terkonsentrasi di kawasan perkotaan besar dan wilayah metropolitan. Ini mencerminkan fakta bahwa adopsi e-commerce cenderung lebih tinggi di wilayah dengan tingkat urbanisasi dan infrastruktur yang lebih maju.

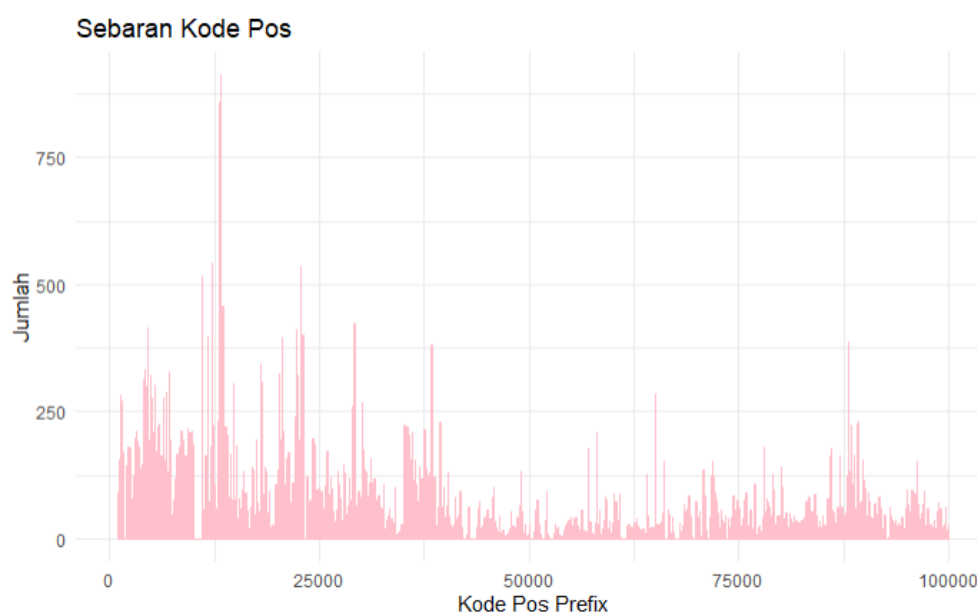
Melalui visualisasi ini, beberapa kesimpulan penting dapat diambil:

- São Paulo adalah pasar terbesar bagi Olist, menunjukkan potensi besar untuk strategi pemasaran, logistik, dan pengembangan layanan khusus di kota tersebut.
- Kota-kota besar lainnya seperti Rio de Janeiro, Belo Horizonte, dan Brasília juga merupakan target penting, meskipun skalanya lebih kecil dibandingkan São Paulo.
- Dengan mengetahui distribusi pelanggan ini, Olist dapat membuat keputusan bisnis yang lebih tepat, seperti mempercepat pengiriman di kota-kota dengan jumlah pelanggan tinggi atau menyesuaikan penawaran produk sesuai preferensi regional.

Grafik ini memberikan gambaran visual yang kuat tentang sebaran pelanggan dan membantu pihak manajemen atau tim analisis dalam merancang strategi bisnis yang lebih fokus berdasarkan wilayah geografis.

## Sebaran Kode Pos

```
```{r}
ggplot(data, aes(x = as.numeric(customer_zip_code_prefix))) +
  geom_histogram(binwidth = 50, fill = "black", color = "pink") +
  labs(title = "Sebaran Kode Pos", x = "Kode Pos Prefix", y = "Jumlah") +
  theme_minimal()
```
```



Hasil visualisasi menunjukkan bahwa distribusi pelanggan tidak merata di seluruh rentang kode pos. Justru terdapat konsentrasi pelanggan yang sangat tinggi pada rentang kode pos di angka-angka yang lebih kecil, terutama antara kode pos 0 hingga sekitar 30.000. Di area ini, banyak sekali lonjakan jumlah pelanggan yang membentuk puncak-puncak tajam, menandakan bahwa ada area tertentu (kode pos tertentu) yang memiliki konsentrasi pelanggan yang jauh lebih besar dibandingkan area lain.

Salah satu puncak tertinggi pada grafik ini terjadi di sekitar kode pos 20.000 hingga 25.000, yang berarti bahwa ada wilayah tertentu dalam rentang tersebut yang menyumbang jumlah pelanggan yang sangat besar bagi Olist. Ini kemungkinan besar berasal dari daerah-daerah padat penduduk seperti São Paulo, Rio de Janeiro, atau kota besar lain di Brasil, mengingat sebelumnya kita sudah melihat São Paulo sebagai kota dengan jumlah pelanggan terbanyak.

Selain itu, terlihat pula bahwa setelah angka 30.000, intensitas pelanggan mulai menurun secara bertahap. Meski tetap ada beberapa lonjakan kecil, tren secara keseluruhan menunjukkan jumlah pelanggan semakin jarang. Pada rentang kode pos di atas 50.000 hingga mendekati 100.000, jumlah pelanggan jauh lebih sedikit dan grafik terlihat semakin rata,

dengan beberapa fluktuasi kecil. Ini menandakan bahwa pelanggan Olist lebih terkonsentrasi di area tertentu dan jauh lebih sedikit di area-area lainnya, mungkin karena faktor seperti kepadatan penduduk, akses internet, atau daya beli masyarakat.

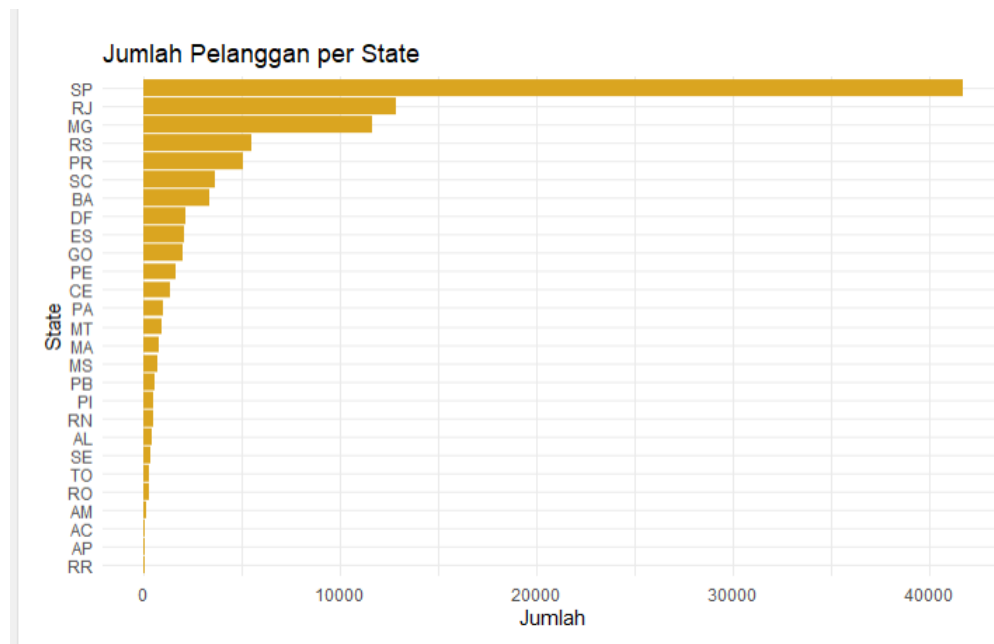
Visualisasi ini memberikan beberapa wawasan penting:

- **Sebaran pelanggan Olist sangat terkonsentrasi di rentang kode pos kecil**, kemungkinan besar di kawasan perkotaan besar dan padat.
- **Area suburban atau rural**, yang biasanya memiliki kode pos lebih tinggi, menyumbang jumlah pelanggan yang relatif kecil.
- **Adanya lonjakan-lonjakan tajam** menunjukkan bahwa wilayah-wilayah tertentu menjadi pusat aktivitas pelanggan yang sangat signifikan dan bisa menjadi target utama untuk strategi pemasaran atau optimalisasi pengiriman logistik.

Secara keseluruhan, histogram ini menegaskan bahwa pelanggan Olist sangat terfokus di beberapa daerah spesifik. Dengan memahami distribusi ini, perusahaan bisa merancang kebijakan yang lebih akurat, misalnya memperbanyak gudang di area-area dengan jumlah pelanggan tertinggi atau menyesuaikan kampanye pemasaran berdasarkan sebaran geografis pelanggan.

## Distribusi Pelanggan per State

```
```{r}
data %>%
  filter(!is.na(customer_state)) %>%
  count(customer_state, sort = TRUE) %>%
  ggplot(aes(x = reorder(customer_state, n), y = n)) +
  geom_col(fill = "goldenrod") +
  coord_flip() +
  labs(title = "Jumlah Pelanggan per State",
       x = "State",
       y = "Jumlah") +
  theme_minimal()
```
```



Dari hasil grafik yang ditampilkan, dapat dilihat dengan sangat jelas bahwa negara bagian São Paulo (SP) menempati posisi teratas dalam jumlah pelanggan untuk perusahaan e-commerce ini, dengan angka yang secara signifikan lebih tinggi dibandingkan negara bagian lainnya. São Paulo, sebagai pusat ekonomi terbesar di Brasil, memang dikenal memiliki infrastruktur bisnis dan teknologi yang sangat maju, yang memungkinkan penetrasi e-commerce berkembang lebih pesat dibandingkan wilayah lain. Tidak hanya itu, tingkat urbanisasi yang tinggi, daya beli masyarakat yang kuat, serta budaya konsumsi yang sudah terintegrasi dengan teknologi digital juga menjadi faktor pendorong dominasi São Paulo dalam grafik tersebut.

Di posisi kedua terdapat Rio de Janeiro (RJ), yang walaupun memiliki jumlah pelanggan yang cukup besar, tetap tertinggal cukup jauh dibandingkan São Paulo. Hal ini mengindikasikan bahwa meskipun Rio de Janeiro juga merupakan kota metropolitan besar dengan potensi ekonomi yang signifikan, faktor-faktor seperti persaingan pasar, tingkat adopsi teknologi, atau bahkan kondisi sosial ekonomi setempat mungkin mempengaruhi skala penggunaan layanan e-commerce. Minas Gerais (MG) menempati posisi ketiga, menunjukkan bahwa negara bagian ini juga memiliki pasar yang cukup potensial, meskipun volumenya masih berada jauh di bawah dua besar.

Setelah ketiga negara bagian utama tersebut, terlihat bahwa jumlah pelanggan dari negara bagian lainnya mulai menurun dengan tajam. Negara bagian seperti Rio Grande do Sul (RS), Paraná (PR), dan Santa Catarina (SC) tetap memberikan kontribusi, namun dalam proporsi yang lebih kecil. Fenomena ini menunjukkan adanya konsentrasi pasar yang berat di wilayah tertentu, sementara di wilayah lain potensi e-commerce masih dalam tahap pertumbuhan atau belum sepenuhnya tergarap secara optimal.

Fenomena distribusi pelanggan yang tidak merata ini dapat memberikan banyak wawasan strategis yang sangat berharga bagi perusahaan e-commerce tersebut. Dengan melihat dominasi São Paulo dan Rio de Janeiro, perusahaan bisa merancang strategi yang lebih terfokus untuk mempertahankan dan memperkuat posisinya di wilayah-wilayah utama tersebut, seperti dengan peningkatan layanan pelanggan, personalisasi pengalaman berbelanja, atau pengembangan infrastruktur logistik yang lebih cepat dan efisien.

Di sisi lain, hasil ini juga membuka peluang bagi perusahaan untuk melakukan ekspansi pasar ke negara-negara bagian yang saat ini jumlah pelanggannya masih rendah, namun mungkin memiliki potensi pertumbuhan yang besar di masa depan. Misalnya, dengan melakukan kampanye pemasaran yang lebih agresif, memberikan insentif khusus untuk pelanggan baru di wilayah-wilayah tersebut, atau bekerja sama dengan mitra lokal untuk meningkatkan brand awareness dan kepercayaan masyarakat terhadap e-commerce.

Selain itu, dalam hal penyediaan layanan logistik, perusahaan dapat mengoptimalkan rute pengiriman dan pengelolaan gudang berdasarkan konsentrasi pelanggan, sehingga dapat menekan biaya operasional sekaligus meningkatkan kecepatan dan kualitas pelayanan. Dari perspektif pengembangan bisnis jangka panjang, analisis semacam ini penting untuk membantu perusahaan mengalokasikan sumber daya secara lebih efektif dan efisien.

Grafik ini, meskipun sederhana, menjadi fondasi awal yang sangat kuat untuk melakukan analisis lanjutan terkait distribusi geografis pelanggan. Dengan pendekatan berbasis data ini, perusahaan bisa merancang perencanaan bisnis yang lebih akurat, memahami dinamika pasar secara lebih baik, serta membuat keputusan strategis yang mendukung pertumbuhan yang berkelanjutan di pasar e-commerce yang semakin kompetitif.