SQL
CAPSTONE PROJECT
ON AMAZON ANALYSIS

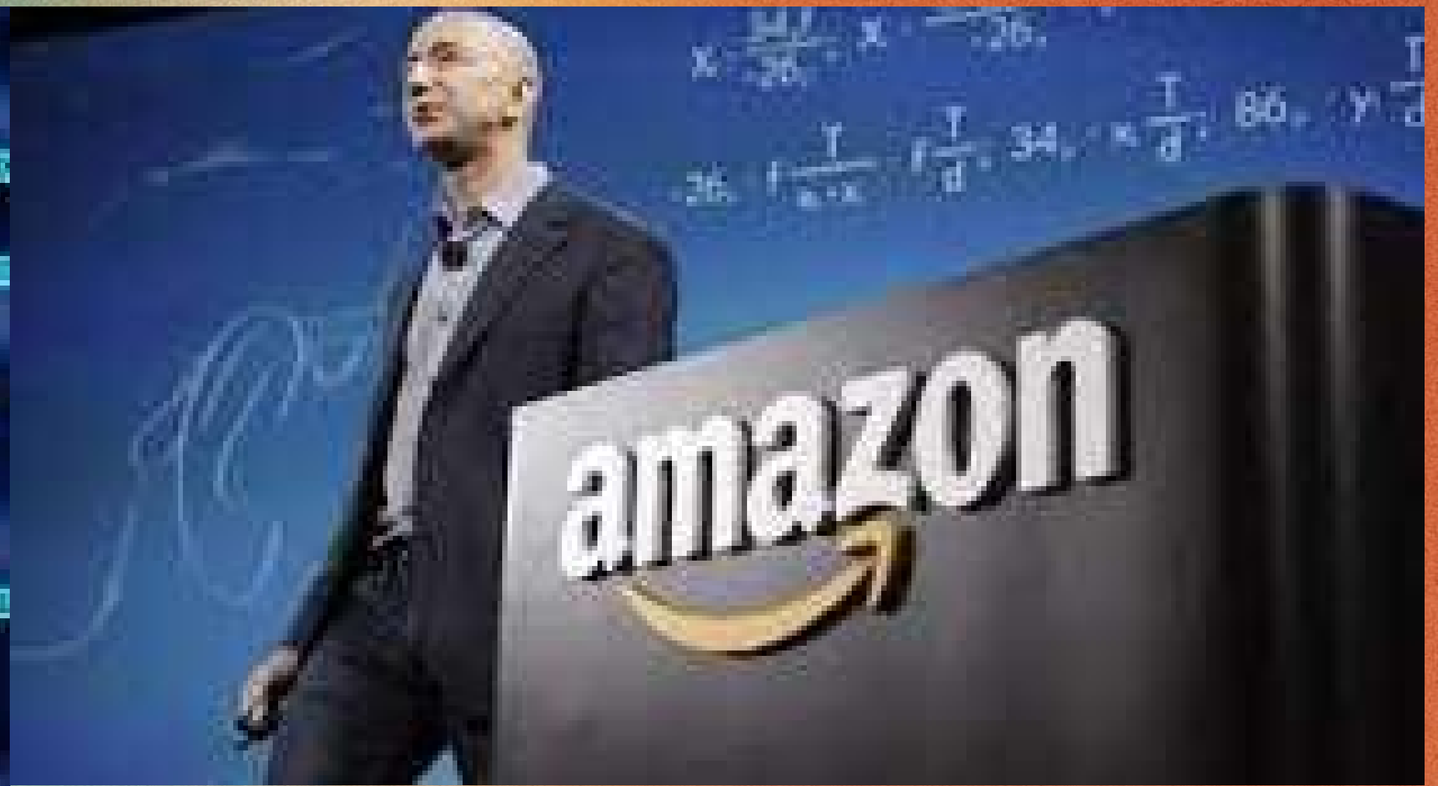


TABLE OF CONTENT

- PRODUCT ANALYSIS
- SALES ANALYSIS
- CUSTOMER ANALYSIS

WELCOME

- **THE MAJOR AIM OF THIS PROJECT IS TO GAIN INSIGHT INTO THE SALES DATA OF AMAZON TO UNDERSTAND THE DIFFERENT FACTORS THAT AFFECT SALES OF THE DIFFERENT BRANCHES.**

FEATURE ENGINEERIG:

- TIMEOFDAY
- DAYNAME
- MONTHNAME

#PRODUCT ANALYSIS

1.. WHAT IS THE COUNT OF DISTINCT PRODUCT LINES IN THE DATASET?

- SELECT COUNT(DISTINCT PRODUCTLINE) AS DISTINCT_PRODUCT_LINES FROM AMAZON;

" 6"

2. WHICH PRODUCT LINE HAS THE HIGHEST SALES?

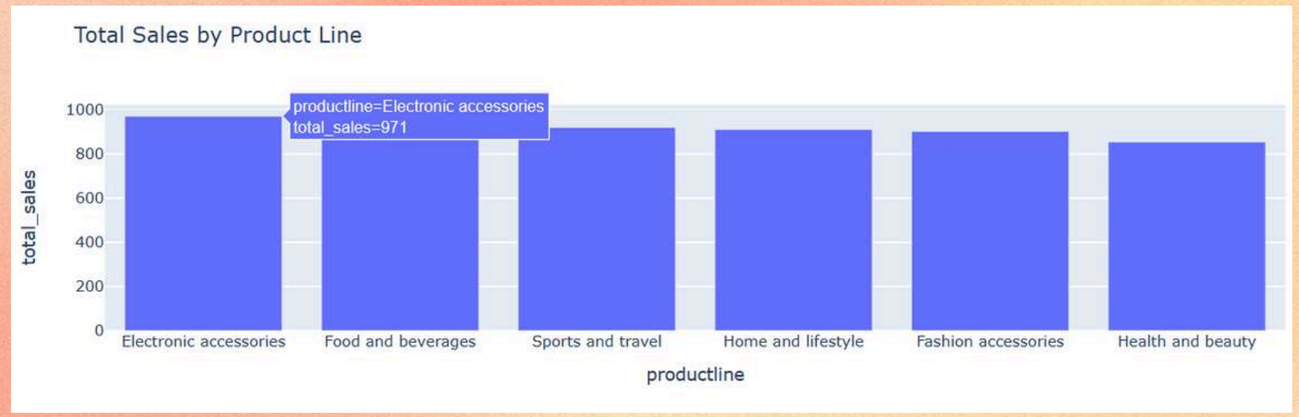
IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- QUERY2 = """ SELECT PRODUCTLINE, SUM(QUANTITY) AS TOTAL_SALES FROM AMAZON GROUP BY PRODUCT_LINE ORDER BY TOTAL_SALES DESC; """

DF2 =PD.READ_SQL (QUERY2,CONN)

FIG2 = PX.BAR(DF2, X='PRODUCTLINE', Y='TOTAL_SALES', TITLE='TOTAL SALES BY PRODUCT LINE'

FIG2.SHOW()



#PRODUCT ANALYSIS

3.WHICH PRODUCT LINE GENERATED THE HIGHEST REVENUE?

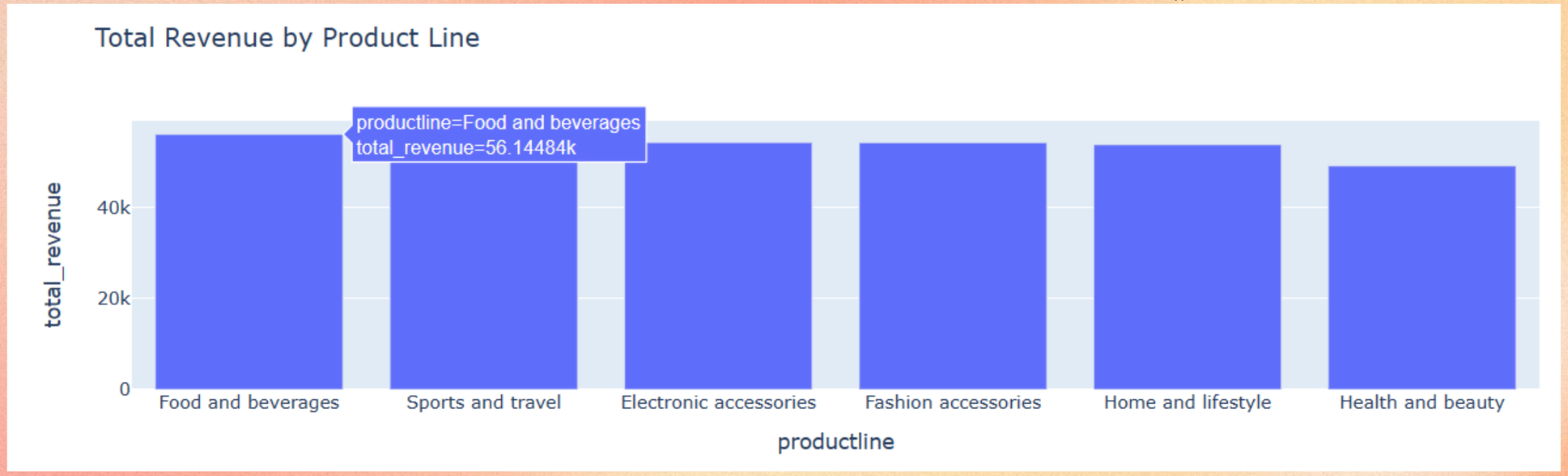
- **QUERY3 = """ SELECT PRODUCTLINE, SUM(TOTAL) AS TOTAL_REVENUE FROM AMAZON GROUP BY PRODUCTLINE ORDER BY TOTAL_REVENUE DESC; """.**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF3 =PD.READ_SQL (QUERY3,CONN)**

FIG3 = PX.BAR(DF3, X='PRODUCTLINE', Y='TOTAL_REVENUE', TITLE='TOTAL REVENUE BY PRODUCT LINE')

- **FIG3.SHOW()**



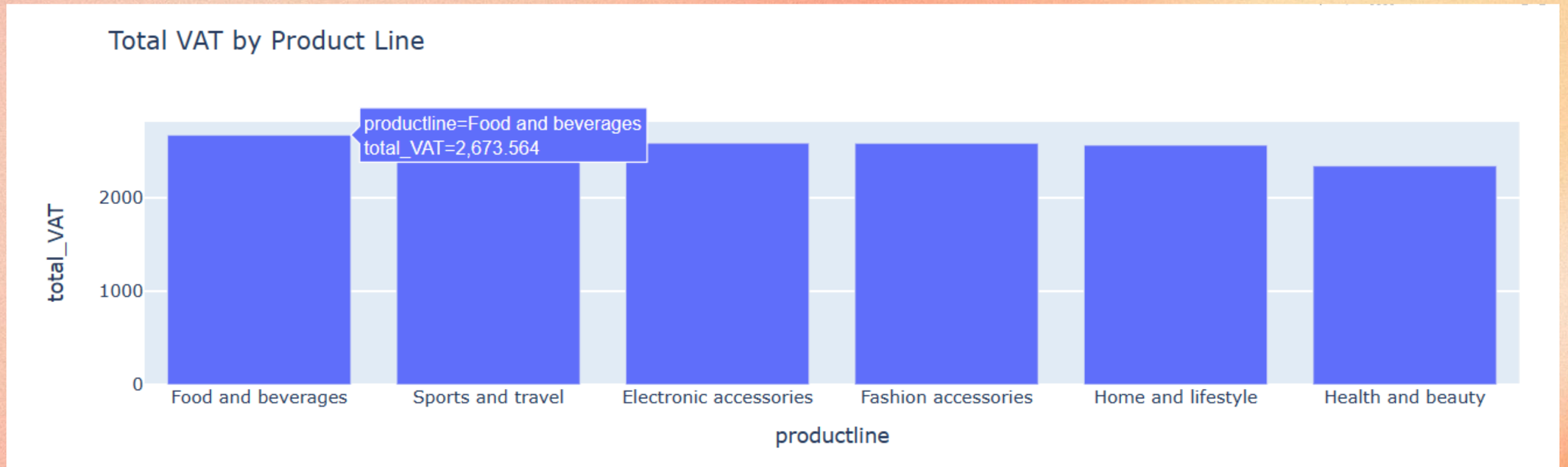
#PRODUCT ANALYSIS

4.WHICH PRODUCT LINE INCURRED THE HIGHEST VALUE ADDED TAX?

- **QUERY4 = """ SELECT PRODUCTLINE, SUM(VAT) AS TOTAL_VAT FROM AMAZON GROUP BY PRODUCTLINE ORDER BY TOTAL_VAT DESC; """**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF4 =PD.READ_SQL (QUERY4,CONN)**
- **FIG4 = PX.BAR(DF4, X='PRODUCTLINE', Y='TOTAL_VAT', TITLE='TOTAL VAT BY PRODUCT LINE')**
- **FIG4.SHOW()**



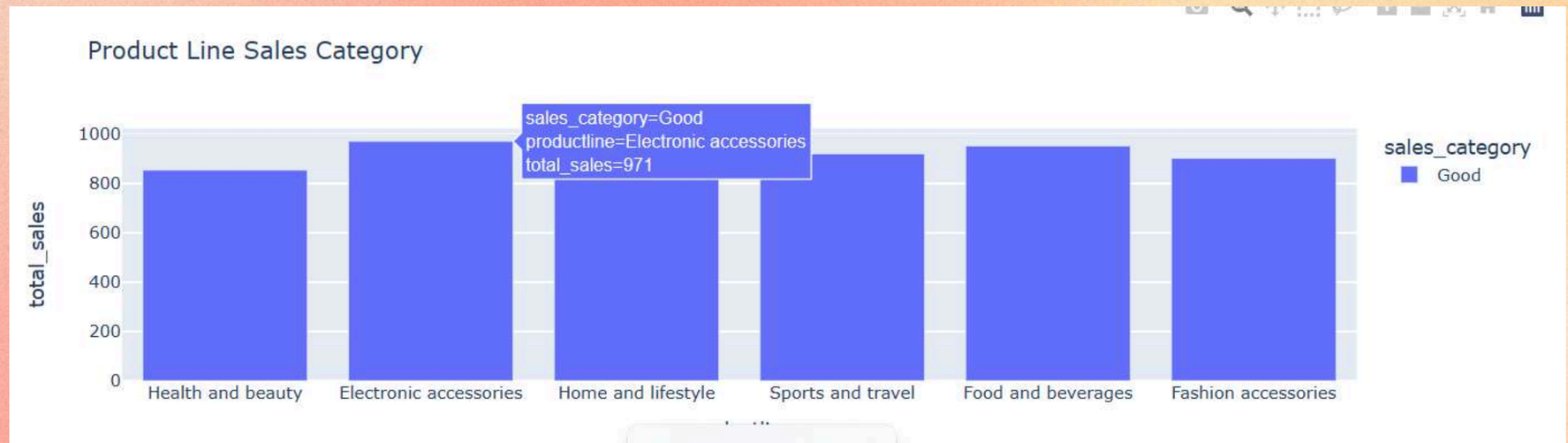
#PRODUCT ANALYSIS

5.FOR EACH PRODUCT LINE, ADD A COLUMN INDICATING "GOOD" IF ITS SALES ARE ABOVE AVERAGE, OTHERWISE "BAD."

```
QUERY5 = """ WITH AVGSALES AS ( SELECT
                AVG(QUANTITY) AS AVG_SALES FROM
AMAZON ) SELECT PRODUCTLINE, SUM(QUANTITY) AS
TOTAL_SALES, CASE WHEN SUM(QUANTITY) > (SELECT
AVG_SALES FROM AVGSALES) THEN 'GOOD' ELSE 'BAD' END
AS SALES_CATEGORY FROM AMAZON GROUP BY
PRODUCTLINE; """
```

```
IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX
```

- `DF5 =PD.READ_SQL (QUERY5,CONN)`
- ```
  FIG5 = PX.BAR(DF5, X='PRODUCTLINE',
  Y='TOTAL_SALES', COLOR='SALES_CATEGORY',
  TITLE='PRODUCT LINE SALES CATEGORY')
  ```
- `FIG5.SHOW()`


  ```

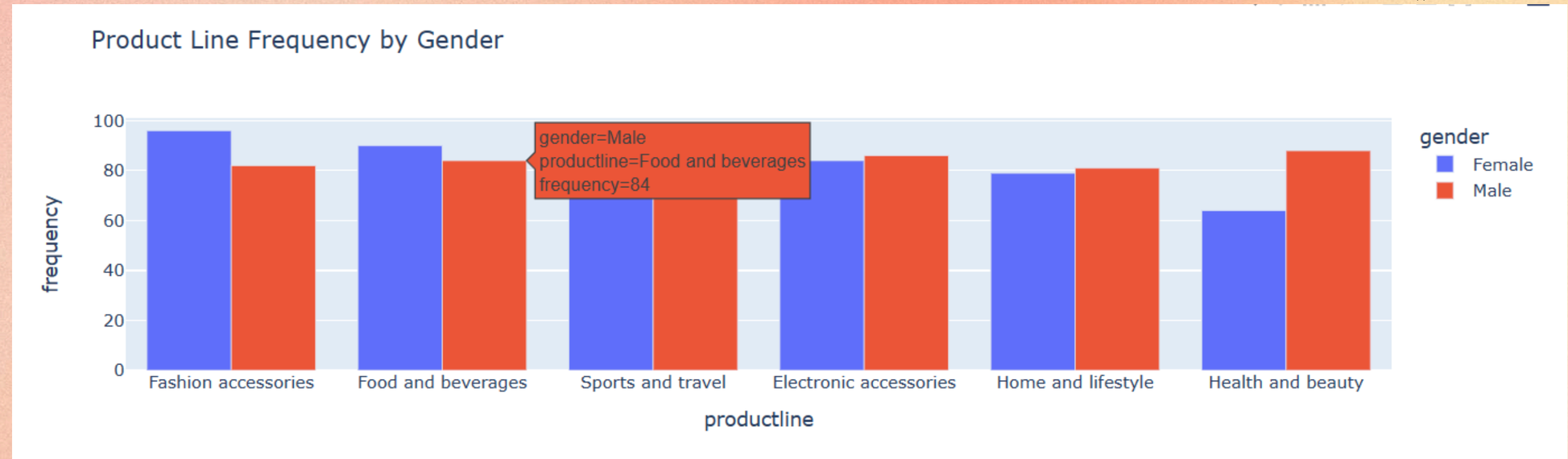
#PRODUCT ANALYSIS

6.WHICH PRODUCT LINE IS MOST FREQUENTLY ASSOCIATED WITH EACH GENDER?

- QUERY6 = """ SELECT GENDER, PRODUCTLINE, COUNT(*) AS FREQUENCY FROM AMAZON GROUP BY GENDER, PRODUCTLINE ORDER BY GENDER, FREQUENCY DESC; """

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF6 =PD.READ_SQL (QUERY6,CONN)
- FIG6 = PX.BAR(DF6, X='PRODUCTLINE', Y='FREQUENCY', COLOR='GENDER',
  TITLE='PRODUCT LINE FREQUENCY BY GENDER',
  BARMODE='GROUP')
- FIG6.SHOW()



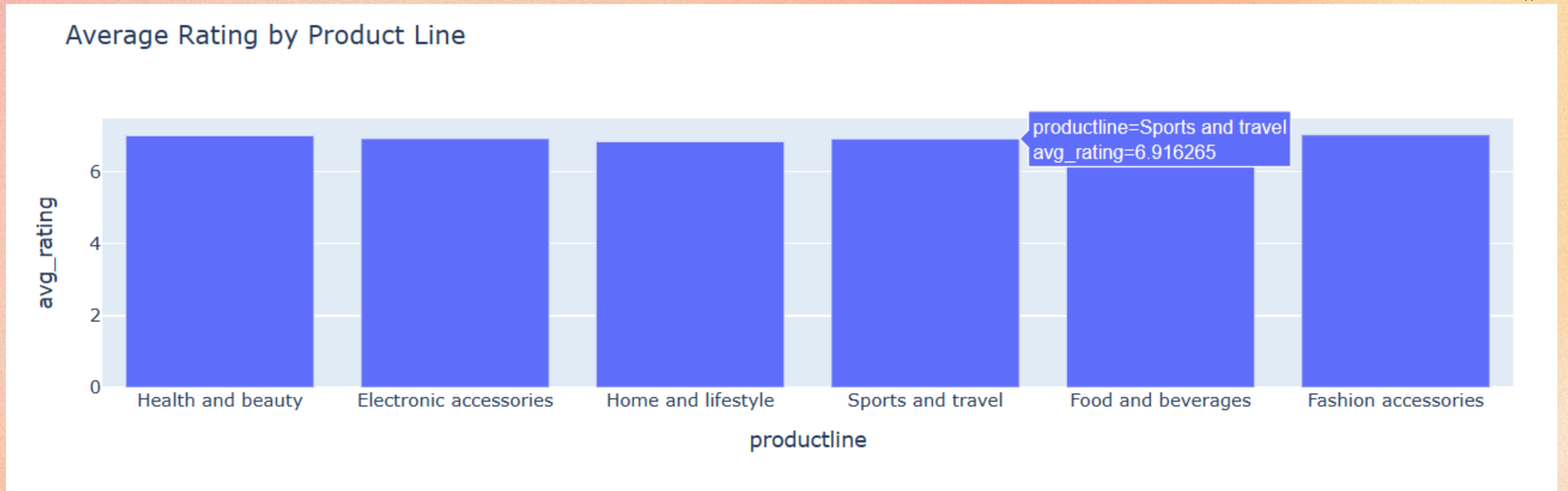
#PRODUCT ANALYSIS

7.CALCULATE THE AVERAGE RATING FOR EACH PRODUCT LINE.

- **QUERY7 = """ SELECT PRODUCTLINE, AVG(RATING) AS AVG_RATING FROM AMAZON GROUP BY PRODUCTLINE; """**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF7 =PD.READ_SQL(QUERY7,CONN)**
- **FIG7 = PX.BAR(DF7, X='PRODUCTLINE', Y='AVG_RATING', TITLE='AVERAGE RATING BY PRODUCTLINE')**
- **FIG7.SHOW()**



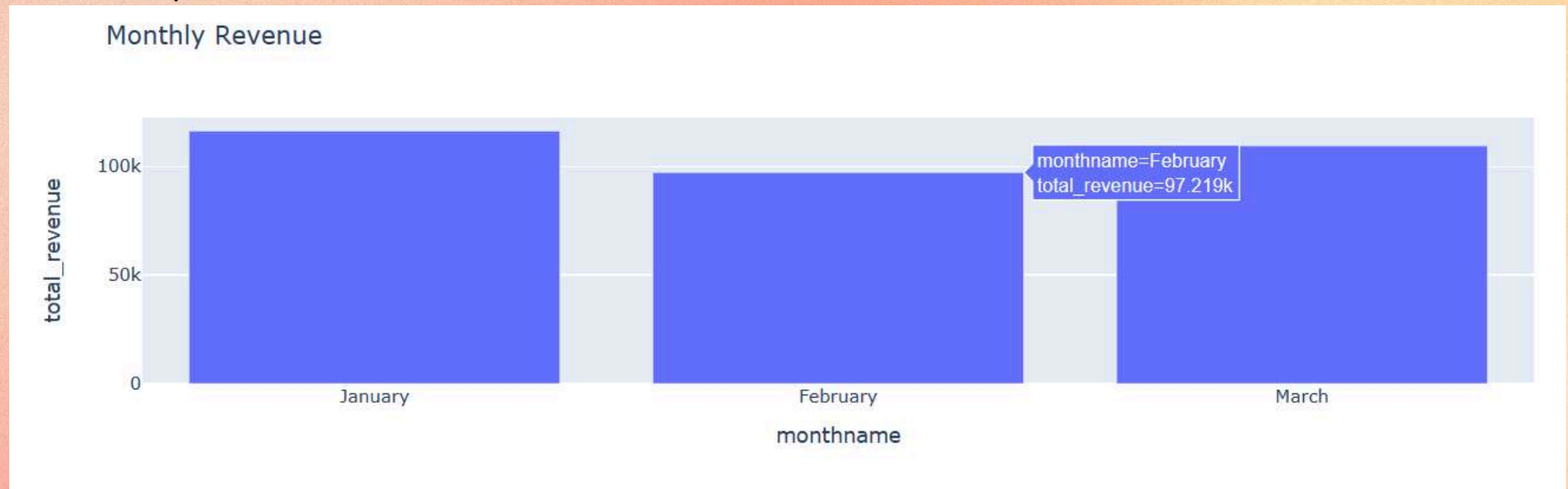
#SALES ANALYSIS

1.HOW MUCH REVENUE IS GENERATED EACH MONTH?

- QUERY8 = """ SELECT MONTHNAME, SUM(TOTAL) AS TOTAL_REVENUE FROM AMAZON
GROUP BY MONTHNAME
ORDER BY FIELD(MONTHNAME, 'JANUARY', 'FEBRUARY', 'MARCH'); """

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF =PD.READ_SQL (QUERY8,CONN)
- FIG = PX.BAR(DATA, X='MONTHNAME', Y='TOTAL_REVENUE', TITLE='MONTHLY REVENUE')
- FIG.SHOW()



#SALES ANALYSIS

2.IN WHICH MONTH DID THE COST OF GOODS SOLD REACH ITS PEAK?

- QUERY9 = """ SELECT MONTHNAME, SUM(COGS) AS TOTAL_COGS FROM AMAZON
  GROUP BY MONTHNAME
  ORDER BY TOTAL_COGS DESC
  LIMIT 1; """

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF =PD.READ_SQL (QUERY9,CONN)
- FIG = PX.BAR(DATA, X='MONTHNAME', Y='TOTAL_COGS', TITLE='MONTH WITH PEAK COGS')
- FIG.SHOW()

#SALES ANALYSIS

3.IDENTIFY THE BRANCH THAT EXCEEDED THE AVERAGE NUMBER OF PRODUCTS SOLD.

- **QUERY10 = """ SELECT BRANCH, SUM(QUANTITY) AS TOTAL_QUANTITY FROM AMAZON GROUP BY BRANCH """**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL(QUERY10,CONN)**
- **FIG = PX.BAR(DATA, X='BRANCH', Y='TOTAL_QUANTITY', TITLE='BRANCHES EXCEEDING AVERAGE QUANTITY SOLD')**
- **FIG.SHOW()**

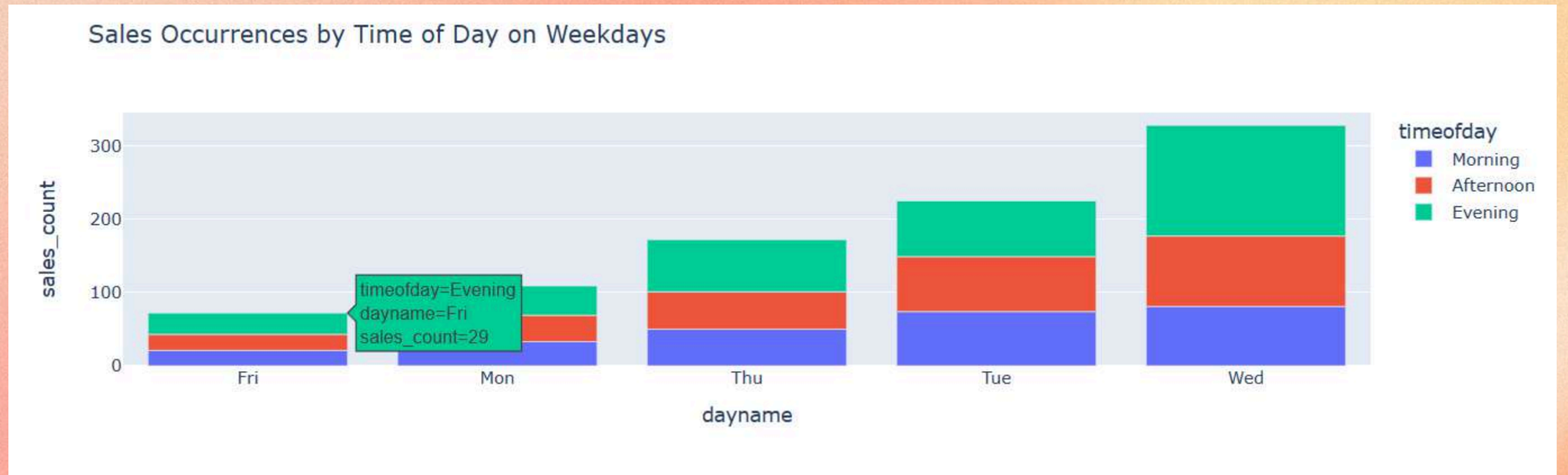
#SALES ANALYSIS

4.COUNT THE SALES OCCURRENCES FOR EACH TIME OF DAY ON EVERY WEEKDAY.

- QUERY11 = """ SELECT DAYNAME, TIMEOFDAY, COUNT(*)
AS SALES_COUNT
FROM AMAZON
GROUP BY DAYNAME, TIMEOFDAY
ORDER BY FIELD(DAYNAME, 'MONDAY', 'TUESDAY',
'WEDNESDAY', 'THURSDAY', 'FRIDAY'),
FIELD(TIMEOFDAY, 'MORNING', 'AFTERNOON',
'EVENING'); """

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF =PD.READ_SQL (QUERY11,CONN)
- FIG = PX.BAR(DATA, X='DAYNAME', Y='SALES_COUNT',
COLOR='TIMEOFDAY', TITLE='SALES OCCURRENCES BY
TIME OF DAY ON WEEKDAYS')
- FIG.SHOW()



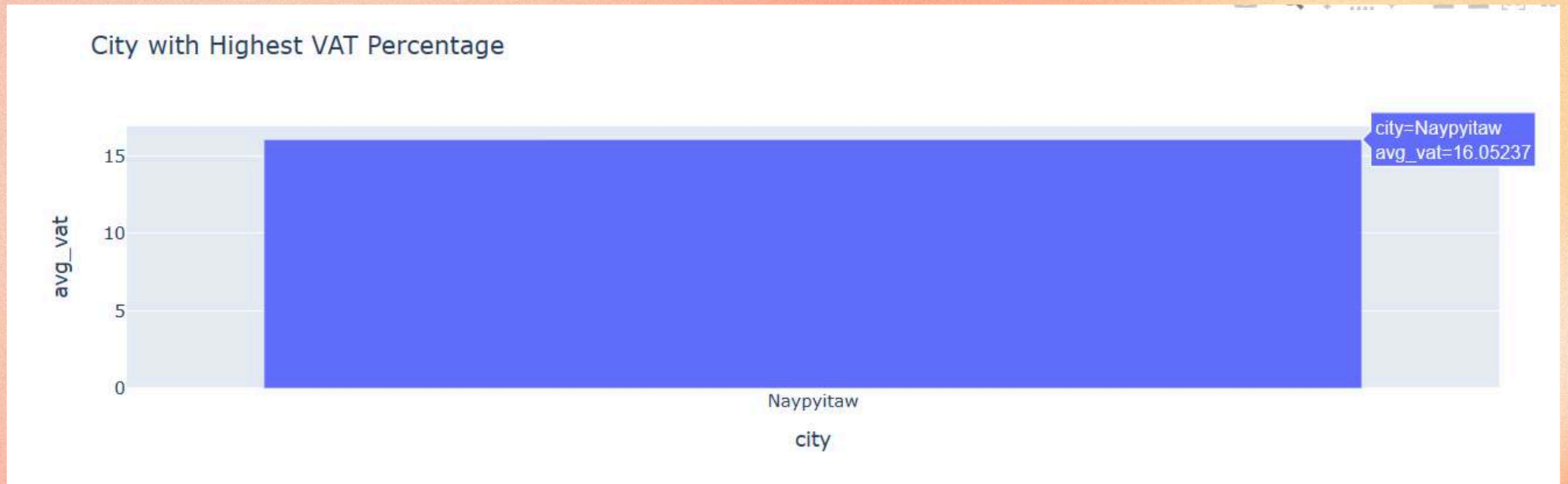
#SALES ANALYSIS

5.DETERMINE THE CITY WITH THE HIGHEST VAT PERCENTAGE.

- **QUERY12 = """ SELECT CITY, AVG(VAT) AS AVG_VAT**
  FROM AMAZON
  GROUP BY CITY
  ORDER BY AVG_VAT DESC
  LIMIT 1; """

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY12,CONN)**
- **FIG = PX.BAR(DATA, X='CITY', Y='AVG_VAT', TITLE='CITY WITH HIGHEST VAT PERCENTAGE')**
- **FIG.SHOW()**



#SALES ANALYSIS

6.WHAT IS THE COUNT OF DISTINCT PAYMENT METHODS IN THE DATASET?

- **QUERY13 = """ SELECT COUNT(DISTINCT PAYMENT) AS DISTINCT_PAYMENT FROM AMAZON; """**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY13,CONN)**
- **FIG = PX.PIE(DATA, NAMES=['PAYMENT METHODS'], VALUES='DISTINCT_PAYMENT_METHODS', TITLE='COUNT OF DISTINCT PAYMENT METHODS')**
- **FIG.SHOW()**

#SALES ANALYSIS

7.WHICH PAYMENT METHOD OCCURS MOST FREQUENTLY?

- QUERY14 = """ SELECT PAYMENT, COUNT(*) AS FREQUENCY
FROM AMAZON
GROUP BY PAYMENT
ORDER BY FREQUENCY DESC
LIMIT 1;
"""

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF =PD.READ_CSV (QUERY14,CONN)
- FIG = PX.BAR(DATA, X='PAYMENT_METHOD', Y='FREQUENCY', TITLE='MOST FREQUENT PAYMENT METHOD')
- FIG.SHOW()

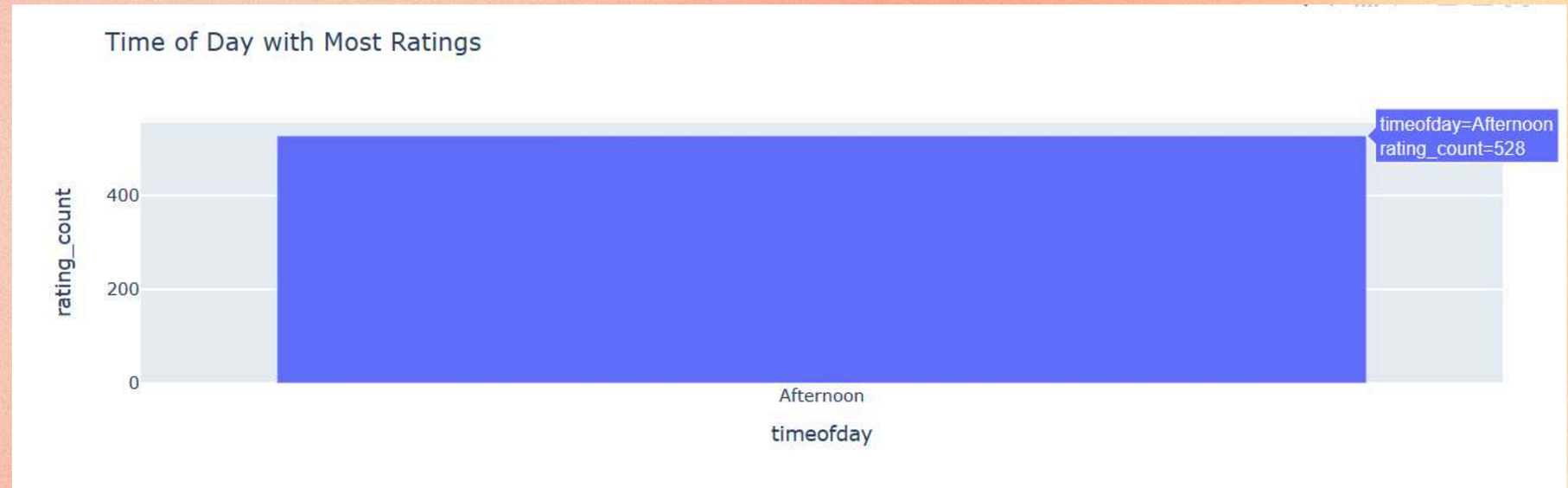
#SALES ANALYSIS

8.IDENTIFY THE TIME OF DAY WHEN CUSTOMERS PROVIDE THE MOST RATINGS.

- **QUERY15 = """ SELECT TIMEOFDAY, COUNT(RATING) AS RATING_COUNT FROM AMAZON GROUP BY TIMEOFDAY ORDER BY RATING_COUNT DESC LIMIT 1;**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY15,CONN)**
- **FIG = PX.BAR(DATA, X='TIMEOFDAY', Y='RATING_COUNT', TITLE='TIME OF DAY WITH MOST RATINGS')**
- **FIG.SHOW()**



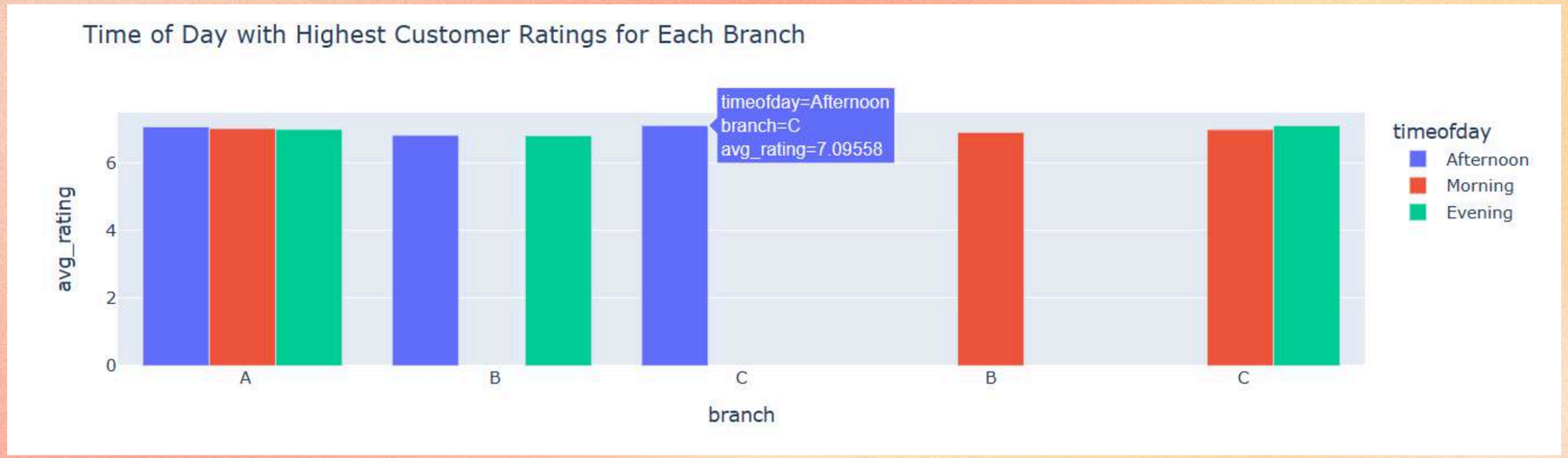
#SALES ANALYSIS

9.DETERMINE THE TIME OF DAY WITH THE HIGHEST CUSTOMER RATINGS FOR EACH BRANCH.

- QUERY16 = """ SELECT BRANCH, TIMEOFDAY,
AVG(RATING) AS AVG_RATING
FROM AMAZON
GROUP BY BRANCH, TIMEOFDAY
ORDER BY BRANCH, AVG_RATING DESC;""""

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF =PD.READ_SQL (QUERY16,CONN)
- FIG = PX.BAR(DATA, X='BRANCH', Y='AVG_RATING', COLOR='TIMEOFDAY', BARMODE='GROUP', TITLE='TIME OF DAY WITH HIGHEST CUSTOMER RATINGS FOR EACH BRANCH')
- FIG.SHOW()



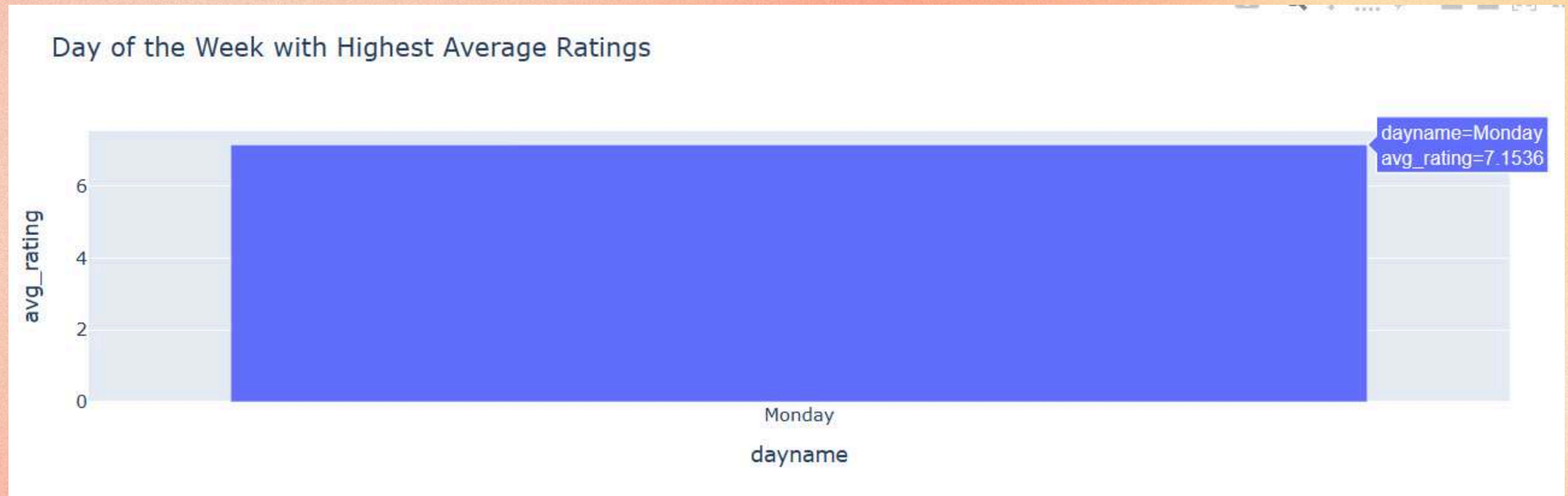
#SALES ANALYSIS

10.IDENTIFY THE DAY OF THE WEEK WITH THE HIGHEST AVERAGE RATINGS.

- QUERY17 = """ SELECT DAYNAME, AVG(RATING) AS AVG_RATING
FROM AMAZON
GROUP BY DAYNAME
ORDER BY AVG_RATING DESC
LIMIT 1;"""""

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF =PD.READ_SQL (QUERY17,CONN)
- FIG = PX.BAR(DATA, X='DAYNAME', Y='AVG_RATING', TITLE='DAY OF THE WEEK WITH HIGHEST AVERAGE RATINGS')
- FIG.SHOW()



#SALES ANALYSIS

11. DETERMINE THE DAY OF THE WEEK WITH THE HIGHEST AVERAGE RATINGS FOR EACH BRANCH.

- QUERY18 = """ SELECT BRANCH, DAYNAME, AVG(RATING) AS AVG_RATING FROM AMAZON GROUP BY BRANCH, DAYNAME ORDER BY BRANCH, AVG_RATING DESC;"""""

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF =PD.READ_SQL (QUERY18,CONN)
- FIG = PX.BAR(DATA, X='BRANCH', Y='AVG_RATING', COLOR='DAYNAME', BARMODE='GROUP', TITLE='DAY OF THE WEEK WITH HIGHEST AVERAGE RATINGS FOR EACH BRANCH')
- FIG.SHOW()

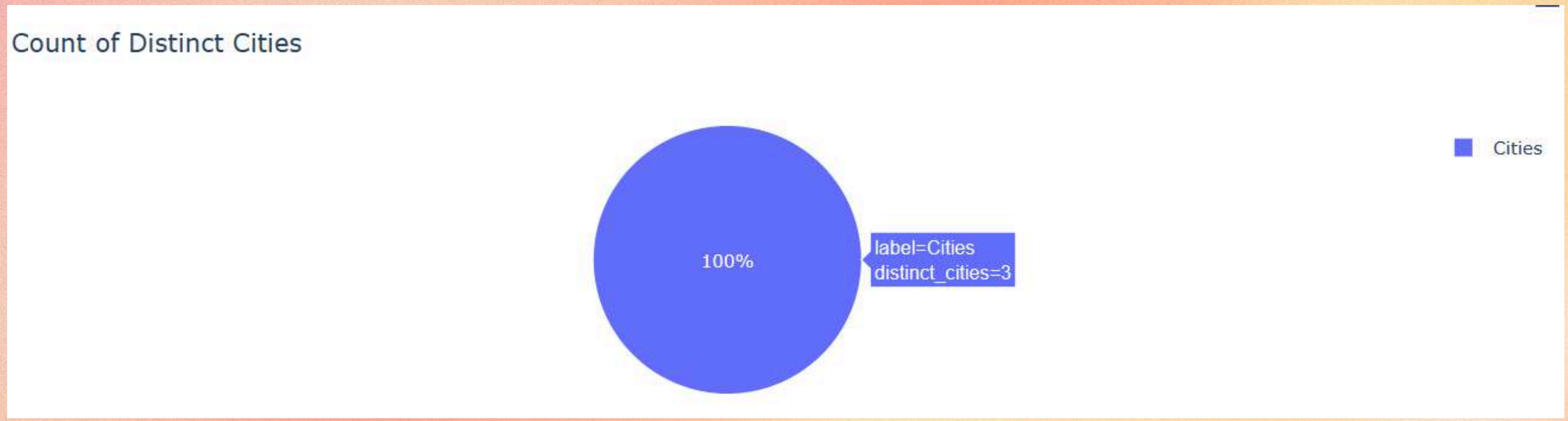
#SALES ANALYSIS

12.WHAT IS THE COUNT OF DISTINCT CITIES IN THE DATASET?

- **QUERY19 = """ SELECT COUNT(DISTINCT CITY) AS DISTINCT_CITIES FROM AMAZON;"""**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY19,CONN)**
- **FIG = PX.PIE(DATA, NAMES=['CITIES'], VALUES='DISTINCT_CITIES', TITLE='COUNT OF DISTINCT CITIES')**
- **FIG.SHOW()**



#SALES ANALYSIS

13.FOR EACH BRANCH, WHAT IS THE CORRESPONDING CITY?

- **QUERY20 = """ SELECT BRANCH, CITY FROM AMAZON GROUP BY BRANCH, CITY;""""**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL(QUERY20,CONN)**
- **FIG = PX.BAR(DATA, X='BRANCH', Y='CITY', TITLE='CITY CORRESPONDING TO EACH BRANCH')**
- **FIG.SHOW()**

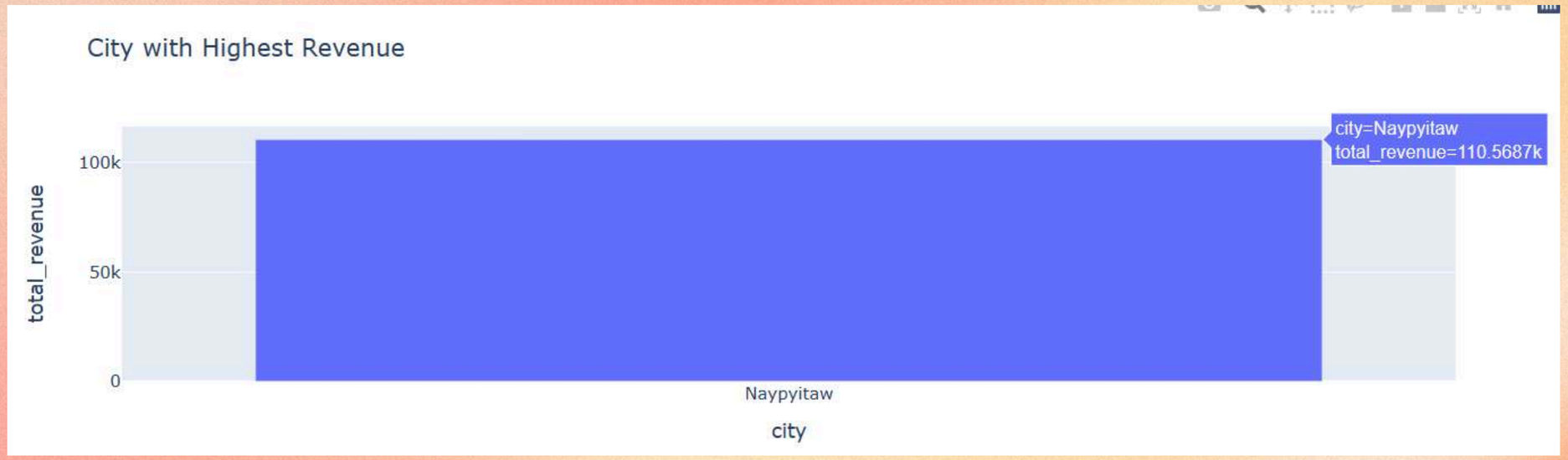
#SALES ANALYSIS

14.IN WHICH CITY WAS THE HIGHEST REVENUE RECORDED?

- **QUERY21 = """ SELECT CITY, SUM(TOTAL) AS TOTAL_REVENUE FROM AMAZON GROUP BY CITY ORDER BY TOTAL_REVENUE DESC LIMIT 1;"""**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY21,CONN)**
- **FIG = PX.BAR(DATA, X='CITY', Y='TOTAL_REVENUE', TITLE='CITY WITH HIGHEST REVENUE')**
- **FIG.SHOW()**



#CUSTOMER ANALYSIS

1. .IDENTIFY THE CUSTOMER TYPE CONTRIBUTING THE HIGHEST REVENUE.

- QUERY22 = """ SELECT CUSTOMERTYPE,
SUM(TOTAL) AS TOTAL_REVENUE
FROM AMAZON
GROUP BY CUSTOMERTYPE
ORDER BY TOTAL_REVENUE DESC
LIMIT 1;
'''''

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF =PD.READ_SQL (QUERY22,CONN)
- FIG = PX.BAR(DF_REVENUE, X='CUSTOMER_TYPE', Y='TOTAL_REVENUE', TITLE='CUSTOMER TYPE WITH HIGHEST REVENUE')
- FIG.SHOW()

#CUSTOMER ANALYSIS

2.IDENTIFY THE CUSTOMER TYPE WITH THE HIGHEST VAT PAYMENTS.

- QUERY23 = """ SELECT CUSTOMERTYPE,
SUM(VAT) AS TOTAL_VAT
FROM AMAZON
GROUP BY CUSTOMERTYPE
ORDER BY TOTAL_VAT DESC
LIMIT 1; """"

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- DF =PD.READ_SQL (QUERY23,CONN)
- FIG = PX.BAR(DF_VAT, X='CUSTOMER_TYPE', Y='TOTAL_VAT', TITLE='CUSTOMER TYPE WITH HIGHEST VAT PAYMENTS')
- FIG.SHOW()

#CUSTOMER ANALYSIS

3.WHAT IS THE COUNT OF DISTINCT CUSTOMER TYPES IN THE DATASET?

- **QUERY24 = """ SELECT COUNT(DISTINCT CUSTOMERTYPE) AS DISTINCT_CUSTOMER_TYPES FROM AMAZON;**
  ,,,,,,

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY24,CONN)**
- **FIG = PX.BAR(DF_DISTINCT, X=DF_DISTINCT.INDEX, Y='DISTINCT_CUSTOMER_TYPES', TITLE='COUNT OF DISTINCT CUSTOMER TYPES')**
- **FIG.SHOW()**

#CUSTOMER ANALYSIS

4.WHICH CUSTOMER TYPE OCCURS MOST FREQUENTLY?

- **QUERY25 = """ SELECT CUSTOMERTYPE,**
  COUNT(*) AS FREQUENCY
  FROM AMAZON
  GROUP BY CUSTOMERTYPE
  ORDER BY FREQUENCY DESC
  LIMIT 1;""""

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY25,CONN)**
- **FIG = PX.BAR(DF_FREQUENCY, X='CUSTOMER_TYPE', Y='FREQUENCY', TITLE='MOST FREQUENT CUSTOMER TYPE')**
- **FIG.SHOW()**

#CUSTOMER ANALYSIS

5.IDENTIFY THE CUSTOMER TYPE WITH THE HIGHEST PURCHASE FREQUENCY.

- **QUERY26 = """ SELECT CUSTOMERTYPE, COUNT(INVOICEID) AS PURCHASE_FREQUENCY FROM AMAZON GROUP BY CUSTOMERTYPE ORDER BY PURCHASE_FREQUENCY DESC LIMIT 1;""""**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY26,CONN)**
- **FIG = PX.BAR(DF_PURCHASE_FREQ, X='CUSTOMER_TYPE', Y='PURCHASE_FREQUENCY', TITLE='CUSTOMER TYPE WITH HIGHEST PURCHASE FREQUENCY')**
- **FIG.SHOW()**

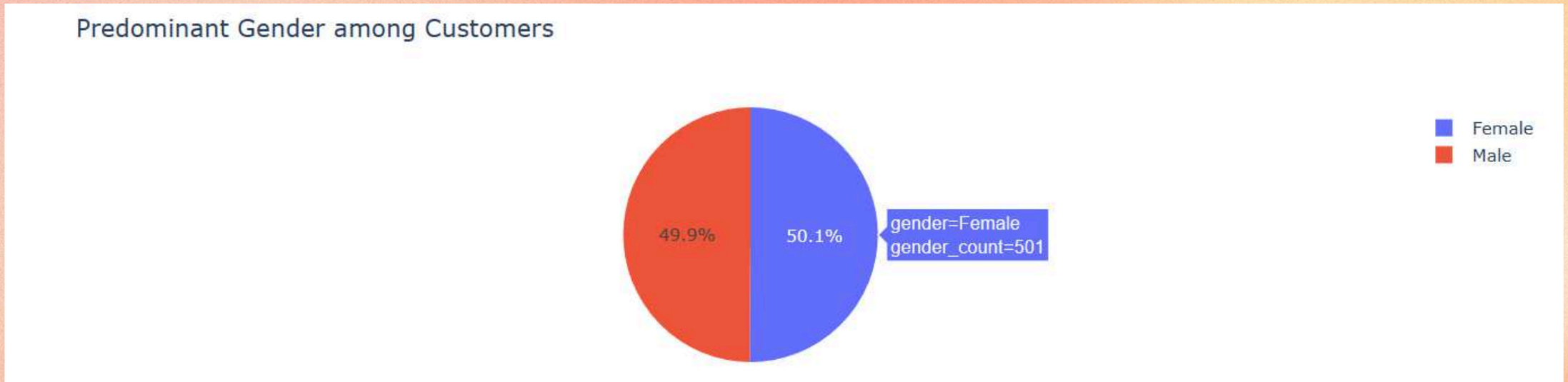
#CUSTOMER ANALYSIS

6.DETERMINE THE PREDOMINANT GENDER AMONG CUSTOMERS.

- **QUERY27 = """ SELECT GENDER, COUNT(*) AS GENDER_COUNT FROM AMAZON GROUP BY GENDER ORDER BY GENDER_COUNT DESC LIMIT 1;""""**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY27,CONN)**
- **FIG = PX.PIE(DF_GENDER, NAMES='GENDER', VALUES='GENDER_COUNT', TITLE='PREDOMINANT GENDER AMONG CUSTOMERS')**
- **FIG.SHOW()**



#CUSTOMER ANALYSIS

7.EXAMINE THE DISTRIBUTION OF GENDERS WITHIN EACH BRANCH.

- **QUERY28 = """ SELECT GENDER, COUNT(*) AS GENDER_COUNT FROM AMAZON GROUP BY GENDER ORDER BY GENDER_COUNT DESC LIMIT 1;""""**

IMPORT PANDAS AS PD
IMPORT PLOTLY.EXPRESS AS PX

- **DF =PD.READ_SQL (QUERY28,CONN)**
- **FIG = PX.BAR(DF_GENDER_BRANCH, X='BRANCH', Y='GENDER_COUNT', COLOR='GENDER', BARMODE='GROUP', TITLE='GENDER DISTRIBUTION')**
- **FIG.SHOW()**

SUMMARY

- High potential in Foods & Beverages product. Increase the sales further by targeting Female category
- Boost Health & Beauty segment by targeting Female also
- Run ads on Home & Lifestyle in the Morning; Sports & Travel in the Afternoon; Food & Beverages in the Evening
- Increase ads and discounts to increase user engagement in the month of February
- Targeted ads and campaign focusing on City and its corresponding gender majority in that city

THANK YOU

PRESENTED BY NALLABOTHULA VENKATESWARLU