

Public Transportation Transaction *

21522747 - Trịnh Tuấn Tú

22520479 - Trương Nguyễn Khánh Hoàng

GVHD: ThS. Đỗ Thị Minh Phụng

Chương trình

- 1.** Giới thiệu đề tài
- 2.** Quá trình xây dựng kho dữ liệu (SSIS)
- 3.** Quá trình phân tích dữ liệu (SSAS)
- 4.** Quá trình lập báo biểu
- 5.** Data Mining
- 6.** Tổng kết



Giới thiệu đề tài

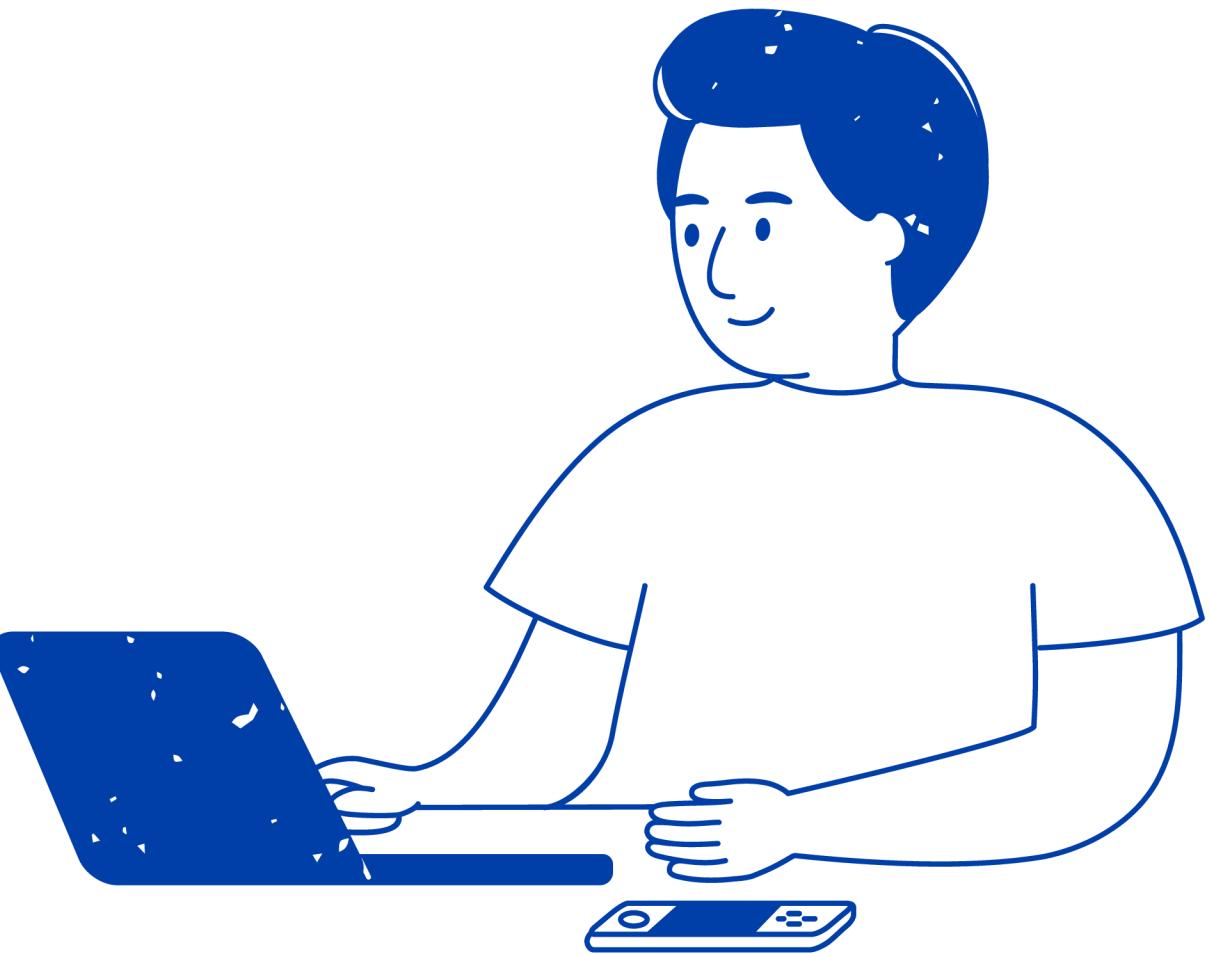
Giới thiệu

Tên bộ dữ liệu: Public Transportation Transaction

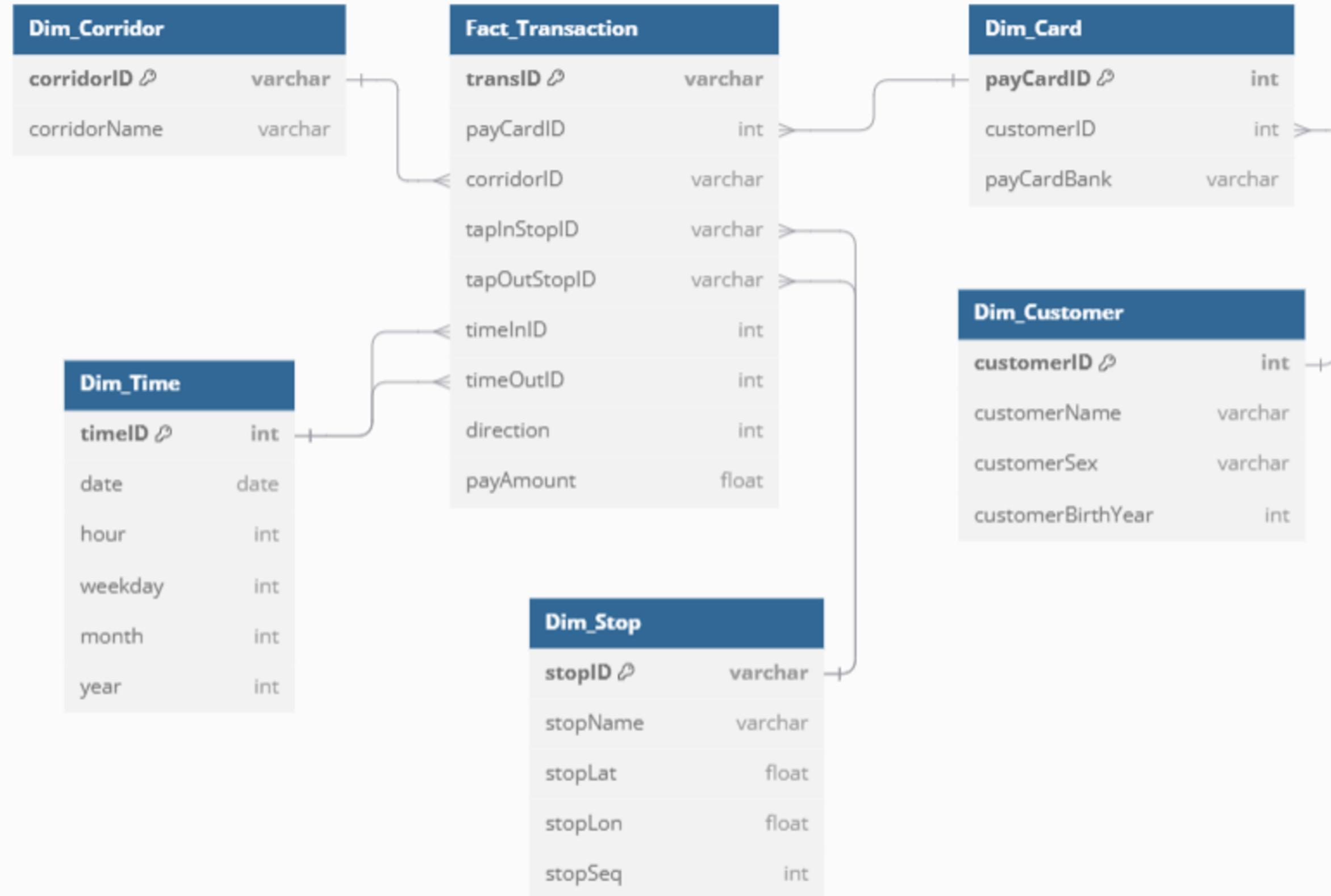
Tác giả: Dikisahkan

Bộ dữ liệu gốc gồm có 2 file csv:

- 1 file 37900 rows, 1 file 189500 rows
- Cả 2 file có cùng số cột là 22 cột



Lược đồ kho dữ liệu



* Tổng quan thuộc tính

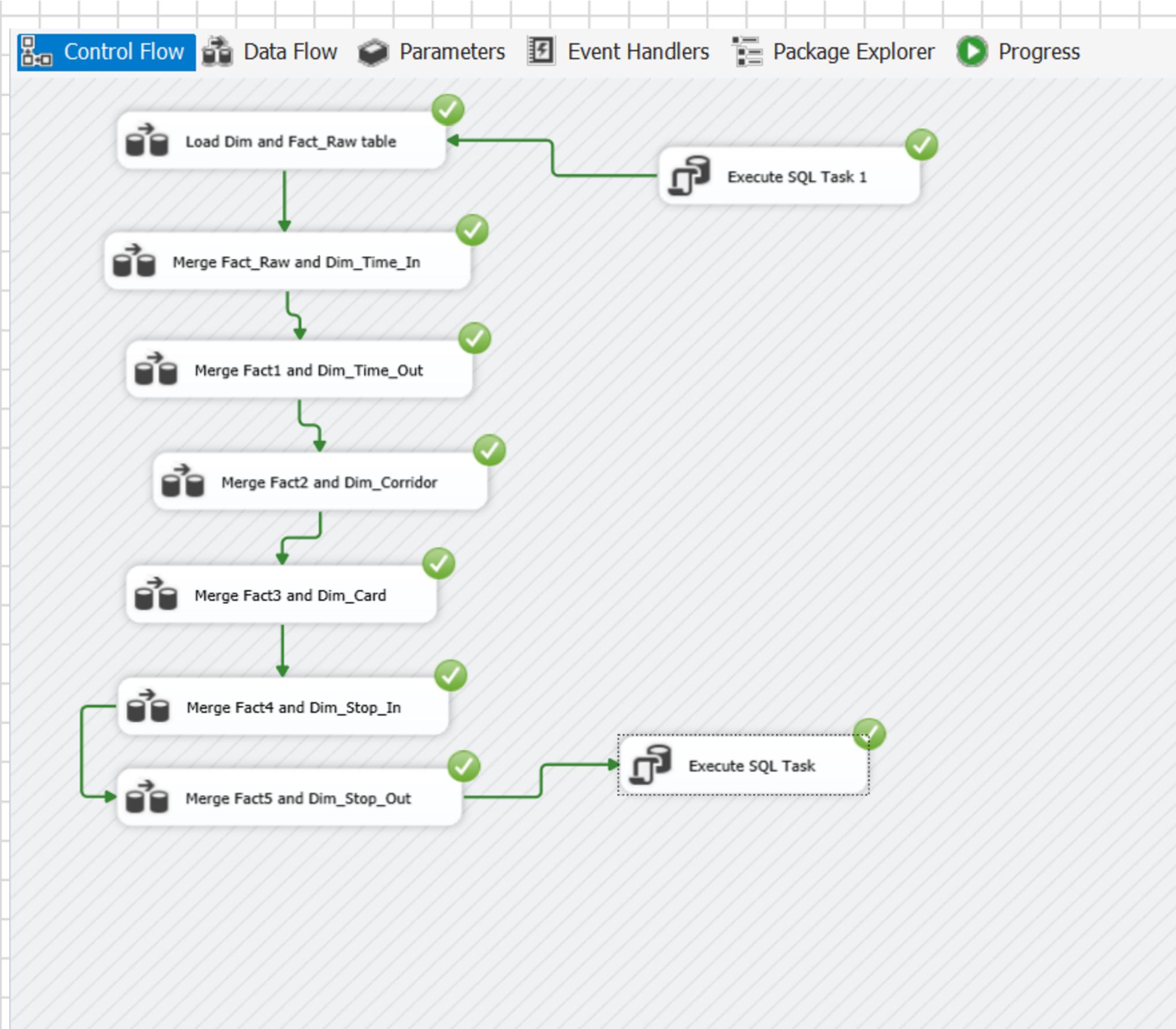
| | | | | |
|----|------------------|--------|----------|---------|
| 0 | transID | 185693 | non-null | object |
| 1 | payCardID | 185693 | non-null | int64 |
| 2 | payCardBank | 185693 | non-null | object |
| 3 | payCardName | 185693 | non-null | object |
| 4 | payCardSex | 185693 | non-null | object |
| 5 | payCardBirthDate | 185693 | non-null | int64 |
| 6 | corridorID | 185693 | non-null | object |
| 7 | corridorName | 185693 | non-null | object |
| 8 | direction | 185693 | non-null | float64 |
| 9 | tapInStops | 185693 | non-null | object |
| 10 | tapInStopsName | 185693 | non-null | object |
| 11 | tapInStopsLat | 185693 | non-null | float64 |

* Tổng quan thuộc tính

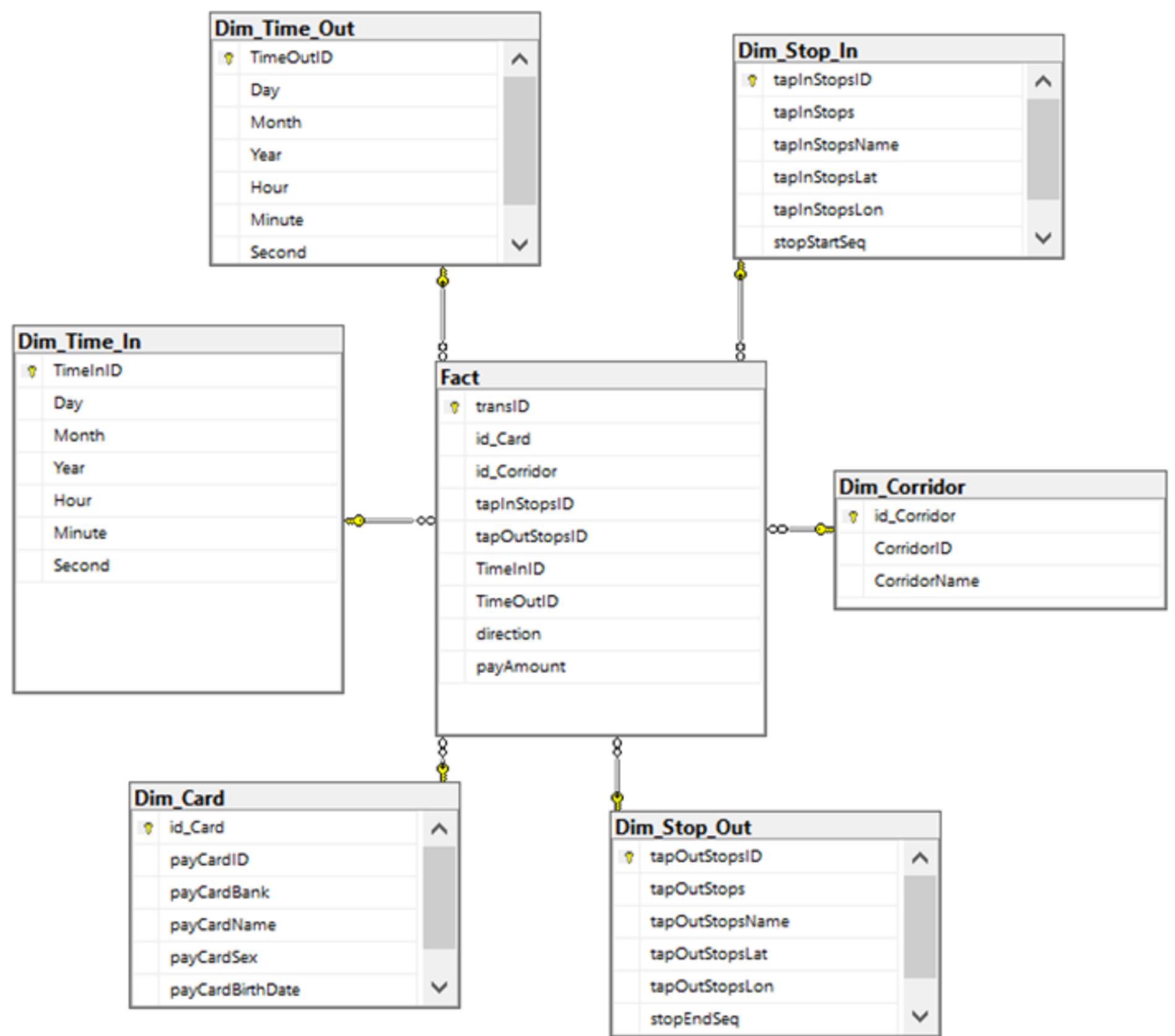
| | | | | |
|----|-----------------|--------|----------|----------------|
| 12 | tapInStopsLon | 185693 | non-null | float64 |
| 13 | stopStartSeq | 185693 | non-null | int64 |
| 14 | tapInTime | 185693 | non-null | datetime64[ns] |
| 15 | tapOutStops | 185693 | non-null | object |
| 16 | tapOutStopsName | 185693 | non-null | object |
| 17 | tapOutStopsLat | 185693 | non-null | float64 |
| 18 | tapOutStopsLon | 185693 | non-null | float64 |
| 19 | stopEndSeq | 185693 | non-null | float64 |
| 20 | tapOutTime | 185693 | non-null | datetime64[ns] |
| 21 | payAmount | 185693 | non-null | float64 |

Quá trình xây dựng kho dữ liệu (SSIS)

Chạy dự án SSIS



Lược đồ sau khi hoàn thành



Quá trình phân tích dữ liệu (SSAS)



Câu truy vấn

1. Top 5 chuyến hành trình có doanh thu thanh toán cao nhất.

```
SELECT  
    ORDER(  
        TOPCOUNT([Dim Corridor].[Corridor Name].children, 5,  
        [Measures].[Total Payment Amount]),  
        [Measures].[Total Payment Amount],  
        DESC) ON ROWS,  
        [Measures].[Total Payment Amount] ON COLUMNS  
FROM [Public Transportation Transaction DW]
```



Câu truy vấn

2. Thống kê tổng số giao dịch và doanh thu theo ngày và giờ.

```
SELECT
    {[Measures].[Fact Count], [Measures].[Total Payment Amount]} ON COLUMNS,
    NONEMPTY(CROSSJOIN(FILTER([Dim Time In].[Day].Members,
        [Dim Time In].[Day].CurrentMember.Name <> "All"
    ),
    FILTER([Dim Time In].[Time In ID].Members,
        [Dim Time In].[Time In ID].CurrentMember.Name <> "All"
    )
)
) ON ROWS
FROM [Public Transportation Transaction DW]
```



Câu truy vấn

3. Tổng doanh thu thanh toán theo trạm đi vào năm 2023.

```
SELECT  
    [Dim Stop In].[Tap In Stops Name].Members ON ROWS,  
    [Measures].[Total Payment Amount] ON COLUMNS  
FROM [Public Transportation Transaction DW]  
WHERE ([Dim Time In].[Year].[2023])
```



Câu truy vấn

4. Top 3 trạm bắt đầu có số lượng giao dịch cao nhất.

```
SELECT  
    TOPCOUNT(  
        FILTER([Dim Stop In].[Tap In Stops Name].Members,  
            [Dim Stop In].[Tap In Stops Name].CurrentMember.Name <> "All"),  
        3,  
        [Measures].[Fact Count]  
    ) ON ROWS,  
    [Measures].[Fact Count] ON COLUMNS  
FROM [Public Transportation Transaction DW]
```



Câu truy vấn

5. Thống kê số tiền khách hàng đã đi theo hướng có Direction.

```
SELECT  
FILTER(  
    [Dim Card].[Pay Card Name].Members,  
    [Measures].[Direction] = 1  
    AND [Measures].[Fact Count] > 2  
) ON ROWS,  
{  
    [Measures].[Direction],  
    [Measures].[Fact Count],  
    [Measures].[Pay Amount]  
} ON COLUMNS  
FROM [Public Transportation Transaction DW]
```



Câu truy vấn

6. Thời gian, số lượng giao dịch, số thứ tự lên và xuống xe ở những chuyến hành trình không có doanh thu.

```
SELECT  
NON EMPTY  
[Measures].[Fact Count] ON COLUMNS,  
  
NON EMPTY  
CROSSJOIN(  
FILTER(  
    [Dim Corridor].[Corridor Name].MEMBERS,  
    [Dim Corridor].[Corridor Name].CURRENTMEMBER.Name <> "All"  
,  
FILTER(  
    [Dim Stop In].[Stop Start Seq].MEMBERS,  
    [Measures].[Total Payment Amount] = 0  
,  
FILTER(  
    [Dim Stop Out].[Stop End Seq].MEMBERS,  
    [Measures].[Total Payment Amount] = 0  
,  
FILTER(  
    [Dim Time In].[Time In ID].MEMBERS,  
    [Dim Time In].[Time In ID].CURRENTMEMBER.Name <> "All"  
)  
)  
DIMENSION PROPERTIES MEMBER_CAPTION, MEMBER_UNIQUE_NAME  
ON ROWS  
  
FROM [Public Transportation Transaction DW]
```

Câu truy vấn

7. Năm sinh nhiều giao dịch nhất trong nhóm tuổi có số lượng giao dịch nhiều nhất

```
SELECT  
    NON EMPTY  
    {[Measures].[Age Group], [Measures].[Fact Count]} ON COLUMNS,  
    NON EMPTY  
    {  
        TOPCOUNT(  
            FILTER(  
                [Dim Card].[Pay Card Birth Date].[Pay Card Birth Date].ALLMEMBERS,  
                [Measures].[Age Group] <> "Unknown"  
            ),  
            1,  
            [Measures].[Fact Count]  
        )  
    } DIMENSION PROPERTIES MEMBER_CAPTION,  
MEMBER_UNIQUE_NAME ON ROWS  
FROM [Public Transportation Transaction DW]  
WHERE  
    ([Dim Time In].[Year].&[2023])
```

Câu truy vấn

8. Thống kê tên chuyến hành trình với số lượng Direction = 1 và Direction = 0 theo ngày trong tháng 4 năm 2023.

```
SELECT  
    NON EMPTY  
    {  
        [Measures].[Query8_1],  
        [Measures].[Query8_2]  
    } ON COLUMNS,  
  
NON EMPTY  
    {  
        ([Dim Time In].[Day].[Day].ALLMEMBERS *  
        [Dim Corridor].[Corridor Name].[Corridor Name].ALLMEMBERS)  
    }  
    DIMENSION PROPERTIES MEMBER_CAPTION,  
    MEMBER_UNIQUE_NAME ON ROWS  
FROM  
(  
    SELECT  
        (  
            {[Dim Time In].[PhanCapTimeIn].[Month].&[4]&[2023]}  
        ) ON COLUMNS  
    FROM  
        [Public Transportation Transaction DW]  
)  
WHERE  
    ([Dim Time In].[PhanCapTimeIn].[Month].&[4]&[2023])
```



Câu truy vấn

9. Danh sách tên, số lượng giao dịch và tổng doanh thu của người đi theo direction bằng 0.

```
SELECT
{
    [Measures].[Fact Count],
    [Measures].[Total Payment Amount]
} ON COLUMNS,
FILTER(
    [Dim Card].[Pay Card Name].Members,
    [Measures].[Direction] = 0
) ON ROWS
FROM [Public Transportation Transaction DW]
```



Câu truy vấn

10. Thống kê số lượng giao dịch của từng vị trí lên xe.

```
SELECT  
{[Measures].[Fact Count]} ON COLUMNS,  
NON EMPTY  
[Dim Stop In].[Tap In Stops Name].MEMBERS ON ROWS  
FROM  
[Public Transportation Transaction DW]
```

Câu truy vấn

11. Top 3 vị trí có số lượng người xuống nhiều nhất.

```
SELECT
{[Measures].[Fact Count]} ON COLUMNS,
NON EMPTY
ORDER(
TOPCOUNT(
FILTER(
[Dim Stop Out].[Tap Out Stops Name].MEMBERS,
[Dim Stop Out].[Tap Out Stops Name].CURRENTMEMBER.Name <>
"All"
),
3,
[Measures].[Fact Count]
),
[Measures].[Fact Count],
DESC
) ON ROWS
FROM
[Public Transportation Transaction DW]
```

Câu truy vấn

12. Điểm đi của ngày nào có số lượng giao dịch cao nhất trong toàn bộ các điểm đi trong tháng 4 năm 2023.

```
SELECT
    {[Measures].[Fact Count]} ON COLUMNS,
    NON EMPTY
TOPCOUNT(

CROSSJOIN(
    FILTER(
        [Dim Stop In].[Tap In Stops Name].MEMBERS,
        [Dim Stop In].[Tap In Stops Name].CURRENTMEMBER.Name <> "All"
    ),
    FILTER(
        [Dim Time In].[Day].MEMBERS,
        [Dim Time In].[Day].CURRENTMEMBER.Name <> "All"
    )
),
1,
[Measures].[Fact Count]
) ON ROWS
FROM
    [Public Transportation Transaction DW]
WHERE
([Dim Time In].[Year].[2023],
[Dim Time In].[Month].[4])
```

Câu truy vấn

13. Thống kê số lượng giao dịch của người dùng theo nhóm tuổi, giới tính, ngân hàng ưa chuộng.

```
SELECT
{
    [Measures].[Age Group],
    [Measures].[Fact Count],
    [Measures].[Total Payment Amount]
} ON COLUMNS,
NONEMPTY(
CROSSJOIN(
    FILTER(
        [Dim Card].[Pay Card Name].MEMBERS,
        [Dim Card].[Pay Card Name].CURRENTMEMBER.Name <> "All"
    ),
    FILTER(
        [Dim Card].[Pay Card Sex].MEMBERS,
        [Dim Card].[Pay Card Sex].CURRENTMEMBER.Name <> "All"
    )
),
FILTER(
    [Dim Card].[Pay Card Bank].MEMBERS,
    [Dim Card].[Pay Card Bank].CURRENTMEMBER.Name <> "All"
),
FILTER(
    [Dim Corridor].[Corridor Name].MEMBERS,
    [Dim Corridor].[Corridor Name].CURRENTMEMBER.Name <> "All"
),
FILTER(
    [Dim Card].[Pay Card Birth Date].[Pay Card Birth
Date].ALLMEMBERS,
    NOT [Measures].[Age Group] = "Unknown"
)
) ON ROWS
FROM [Public Transportation Transaction DW]
```



Câu truy vấn

14. Top 5 năm sinh có số giao dịch cao nhất

```
SELECT  
    NON EMPTY  
    {[Measures].[Age], [Measures].[Age Group], [Measures].[Fact Count]} ON  
    COLUMNS,  
    NON EMPTY  
    { TOPCOUNT(  
        FILTER(  
            [Dim Card].[Pay Card Birth Date].[Pay Card Birth Date].ALLMEMBERS,  
            [Measures].[Age Group] = [Measures].[Age Group]  
        ), 5, [Measures].[Fact Count])  
    } DIMENSION PROPERTIES MEMBER_CAPTION,  
MEMBER_UNIQUE_NAME ON ROWS  
FROM [Public Transportation Transaction DW]
```

*

Câu truy vấn

15. Thống kê các ngân hàng được sử dụng để thanh toán

```
SELECT  
NON EMPTY  
[Dim Card].[Pay Card Bank].MEMBERS ON ROWS,  
NON EMPTY  
[Measures].[Fact Count] ON COLUMNS  
FROM [Public Transportation Transaction DW]
```

Quá trình lập báo biểu

Power BI

Thống kê tổng số giao dịch và doanh thu theo ngày và giờ

Revenue and Transactions

498M

Total Payment Amount

2.98K

Average of Total Payment Amount

186K

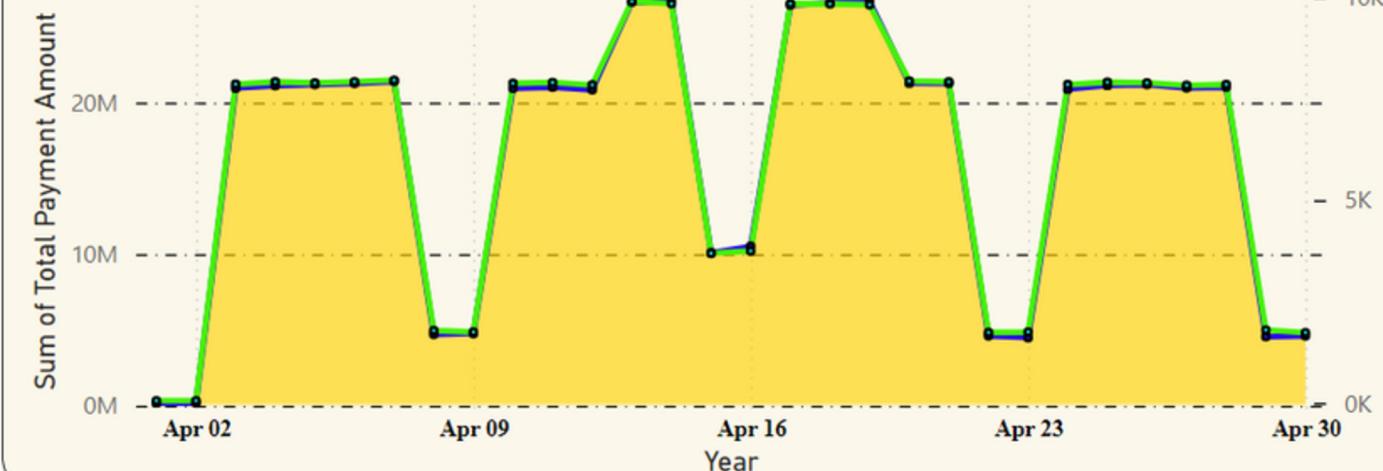
Sum of Number of Transaction

5.71

Standard deviation of Hour

Revenue and Transaction by Month and Year

● Sum of Total Payment Amount ● Sum of Number of Transaction



Year

All

Month

All

Day

All

| Day | 5 | 6 | 7 | 8 | 9 | 10 |
|-------|----------|----------|----------|----------|----------|---------|
| 1 | 10500 | 30500 | 7000 | 0 | 17500 | 3500 |
| 2 | 3500 | 0 | 3500 | | 7000 | 7000 |
| 3 | 1855500 | 3673000 | 1767000 | 1620000 | 1318000 | 3500 |
| 4 | 1977000 | 3707000 | 1584000 | 1655500 | 1726000 | 23500 |
| 5 | 1932500 | 3548500 | 1580500 | 2024500 | 1557000 | 10500 |
| 6 | 1885000 | 3623500 | 1598000 | 1787500 | 1529000 | 3500 |
| 7 | 1917000 | 3529500 | 1638000 | 1702500 | 1733000 | 14000 |
| 8 | 229000 | 293500 | 308000 | 302000 | 244000 | 209000 |
| 9 | 276000 | 305500 | 282000 | 350000 | 284500 | 295000 |
| 10 | 2176500 | 3334500 | 1822000 | 1669500 | 1476000 | 0 |
| 11 | 1904500 | 3583500 | 1639000 | 1709000 | 1629000 | 7000 |
| 12 | 1808000 | 3674500 | 1629500 | 1694500 | 1481500 | 7000 |
| 13 | 2146000 | 4043500 | 1820500 | 1979500 | 2038500 | 289000 |
| 14 | 2138500 | 3929500 | 1860000 | 1848500 | 2072500 | 291000 |
| 15 | 531000 | 568500 | 491500 | 620500 | 610000 | 531000 |
| 16 | 530000 | 581500 | 646000 | 624000 | 662500 | 616500 |
| 17 | 2205000 | 3808500 | 1617500 | 2345000 | 1912000 | 326500 |
| 18 | 2201000 | 4023500 | 1833000 | 1833500 | 2106000 | 248000 |
| 19 | 2304500 | 3832500 | 1831500 | 2005000 | 1996500 | 286500 |
| 20 | 1965500 | 3570000 | 1593500 | 1838500 | 1556500 | 0 |
| Total | 43155500 | 75595500 | 36452000 | 38749500 | 37461500 | 4408000 |

Power BI

Thời gian, số lượng giao dịch, số thứ tự lên và xuống xe ở những chuyến hành trình không có doanh thu



Power BI

Thống kê số lượng giao dịch của người dùng theo nhóm tuổi, giới tính, ngân hàng ưa chuộng.

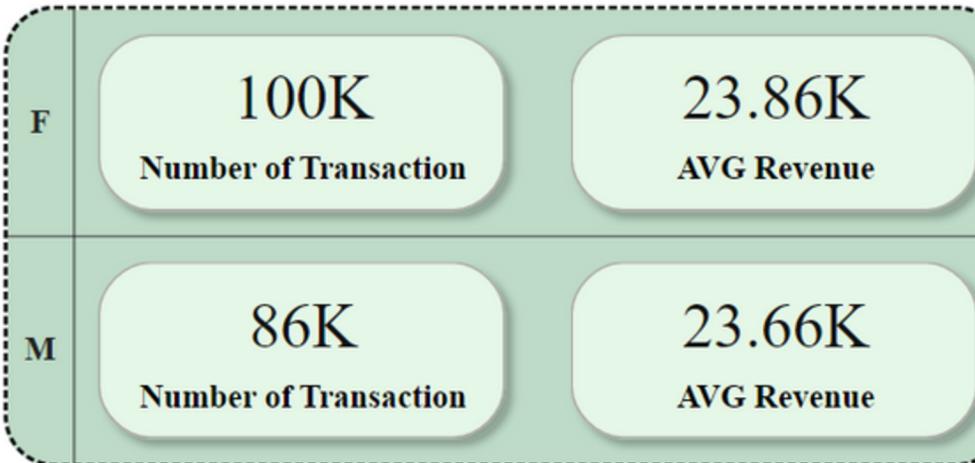
User behavior analysis

Corridor Name
All

Gender
All

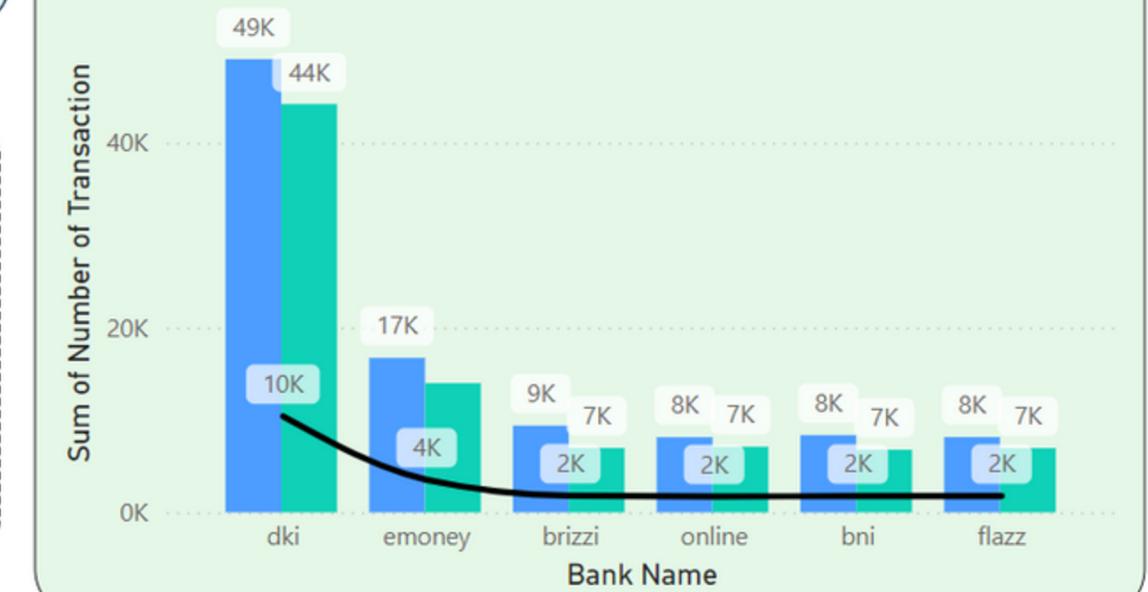
Age Group, Year...
All

Total Payment Amo...
All

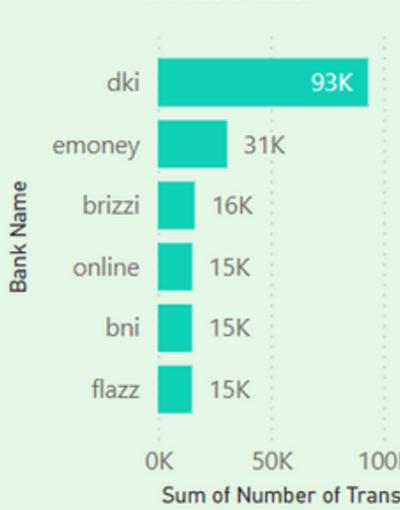


Number of Transaction and Age Group by Bank Name and Gender

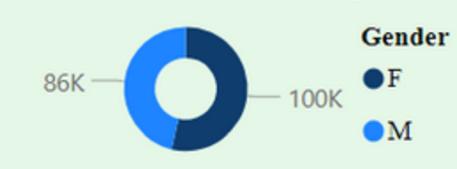
Gender ● F ● M ● Age Group



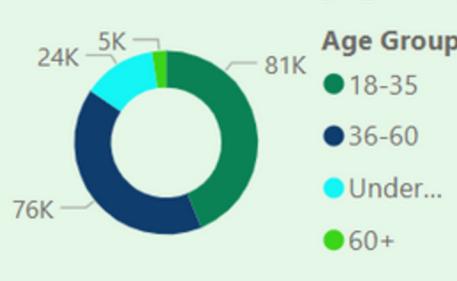
Number of Transaction by Bank Name



Number of Transaction by Gender



Number of Transaction by Age Group

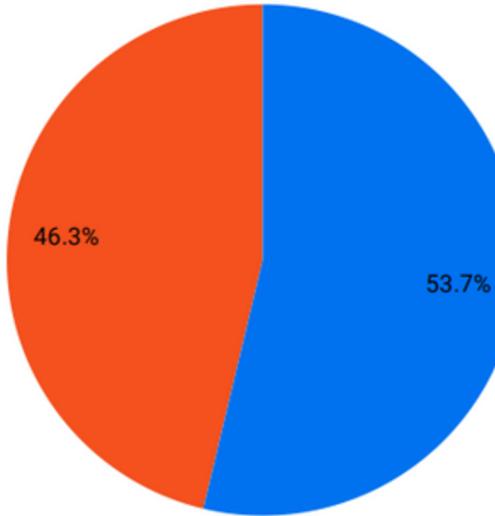


| Name User | Gender | Birth | Bank Name | Corridor Name | Number of Transaction |
|----------------------|--------|-------|-----------|--|-----------------------|
| Zulfa Yuliarti S.Psi | F | 1999 | emoney | BSD - Jelambar | 30 |
| Zulfa Yuliarti | M | 1968 | dki | PGC - Juanda | 1 |
| Zulfa Waskita | F | 1981 | online | Kampung Melayu - Pulo Gebang via BKT | 1 |
| Zulfa Waskita | F | 1981 | online | Kp. Rambutan - Kalisari | 1 |
| Zulfa Waskita | F | 1981 | online | Rusun Jatinegara Kaum - Pulo Gadung | 1 |
| Zulfa Waskita | F | 1981 | online | Senen - Pisangan Baru | 1 |
| Zulfa Waluyo S.Psi | M | 1989 | dki | Gondangdia - Senen | 10 |
| Zulfa Wahyudin | M | 1966 | dki | Puri Beta - Ragunan | 32 |
| Zulfa Wacana | F | 1987 | flazz | Grogol - Strengseng | 1 |
| Zulfa Wacana | F | 1987 | flazz | Harapan Baru - Pulo Gebang via Rawa Kuning | 1 |

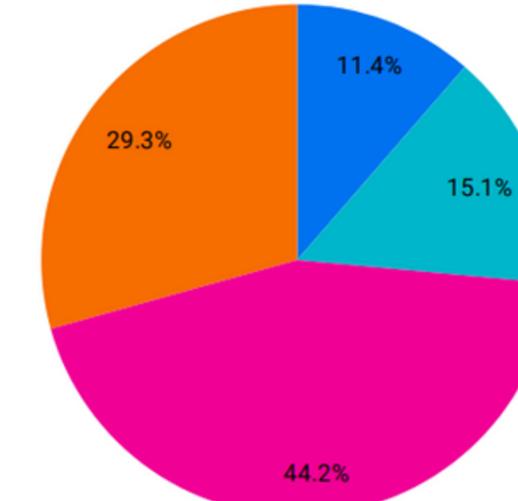
Looker

Doanh thu theo giới tính và độ tuổi.

Giới tính



Độ tuổi



Bảng thống kê doanh thu và số lượng giao dịch theo giới tính và độ tuổi

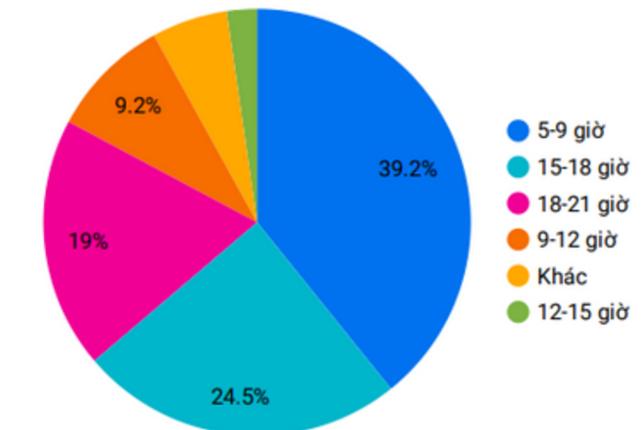
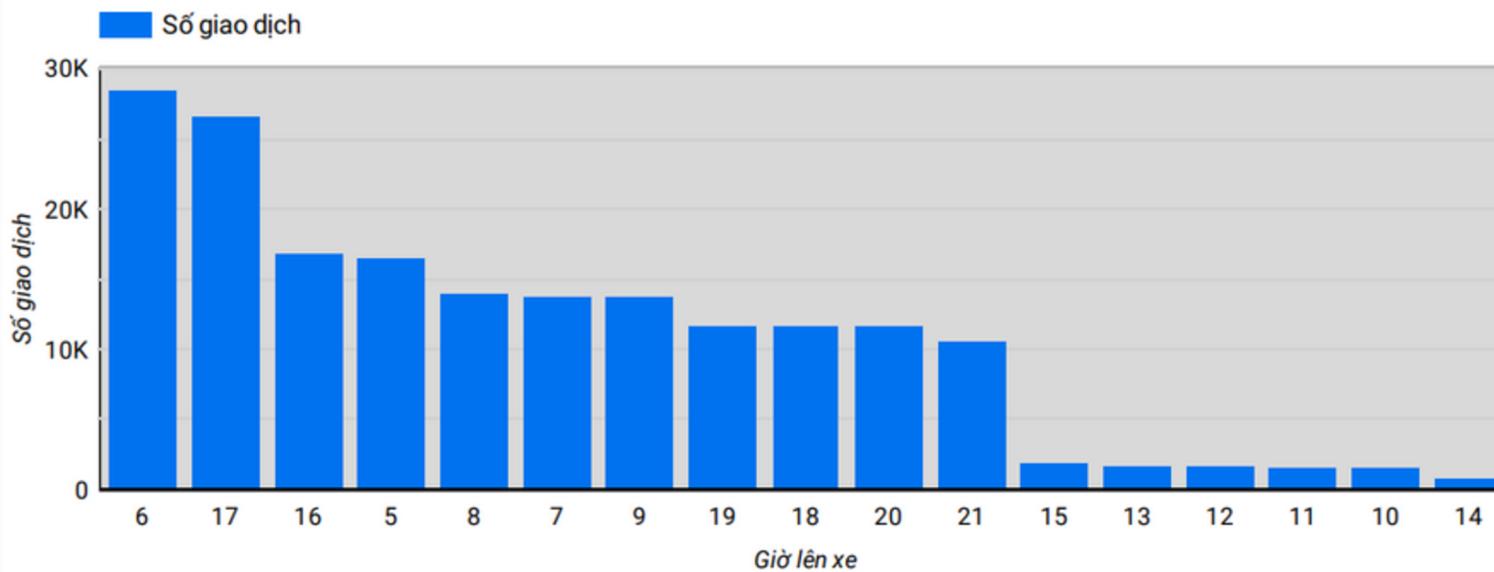
| Giới tính | Age Group / Doanh thu / Số giao dịch | | | | | | | | |
|-------------|--------------------------------------|--------------|-----------|--------------|-----------|--------------|-----------|--------------|-----------|
| | 31-50 | | 18-30 | | Dưới 18 | | Trên 50 | | |
| | Doanh thu | Số giao dịch | Doanh thu | Số giao dịch | Doanh thu | Số giao dịch | Doanh thu | Số giao dịch | Doanh thu |
| F | 118.7M | 43.5K | 93M | 33.8K | 43.4M | 17.1K | 11.9M | 5.4K | |
| M | 104.9M | 38.5K | 56.4M | 20.7K | 28.9M | 11K | 41M | 15.7K | |
| Grand total | 223.6M | 82.1K | 149.4M | 54.5K | 72.4M | 28K | 52.9M | 21.1K | |

Looker

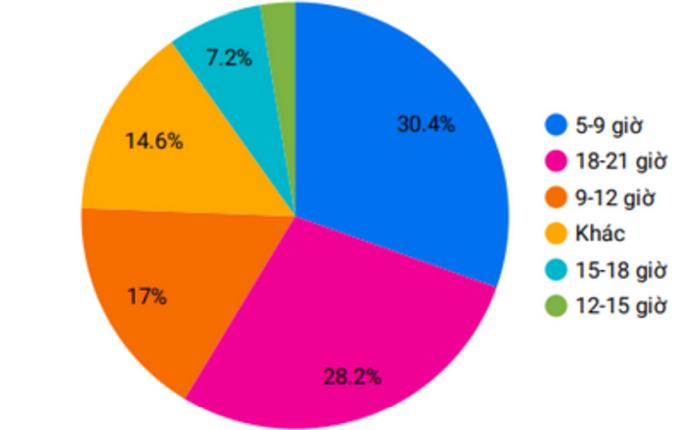
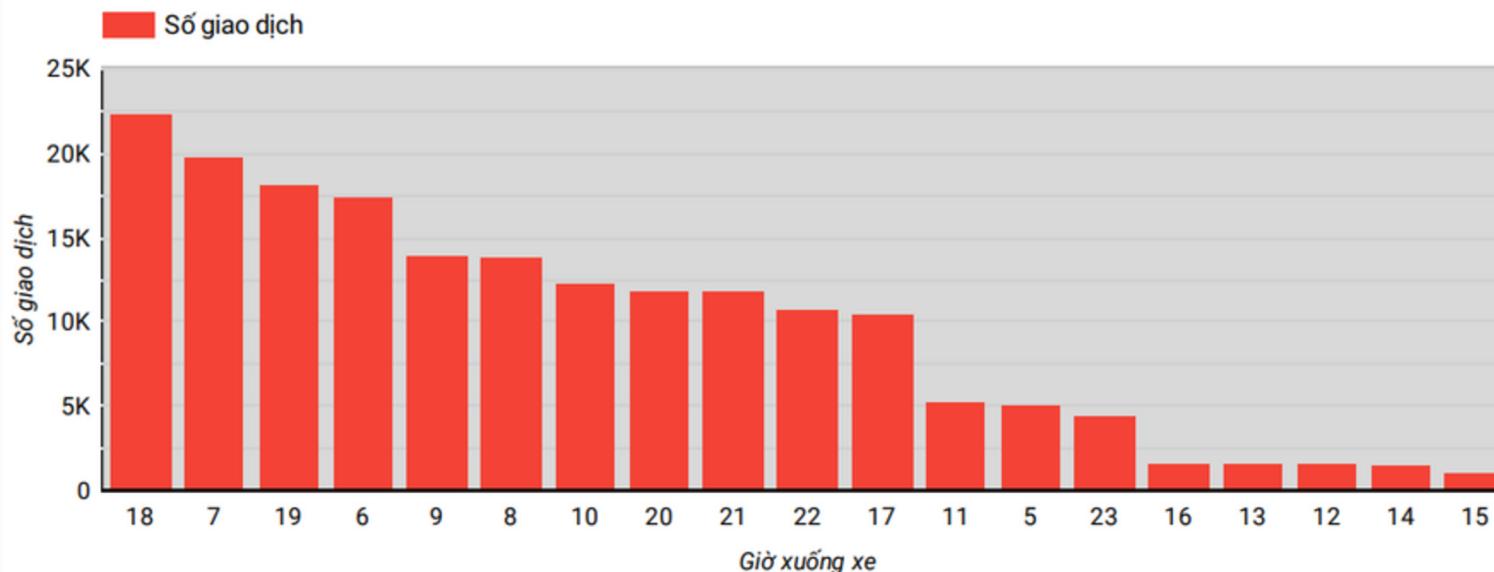
Thống kê phân tích giờ cao điểm.

Số giao dịch
185.7K

Thời gian hành khách lên xe



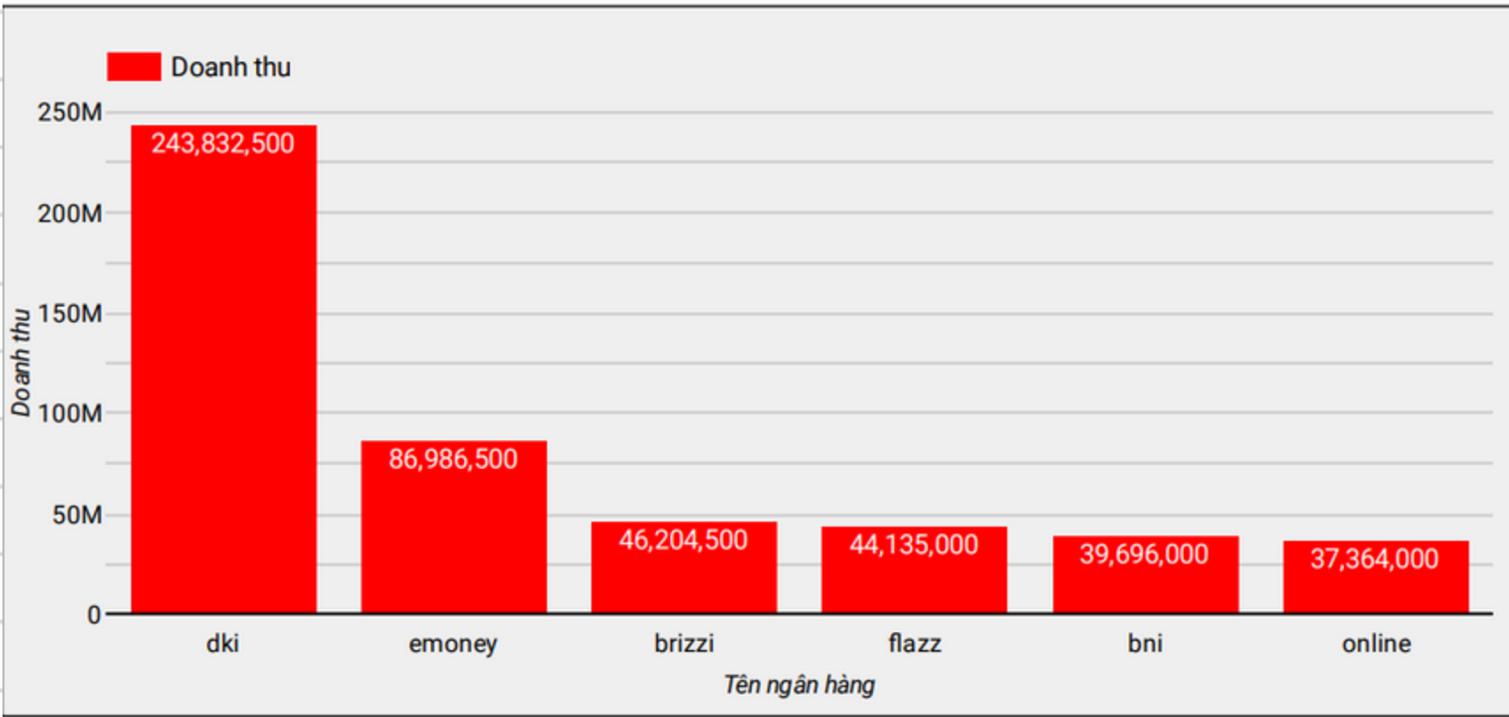
Thời gian hành khách xuống xe



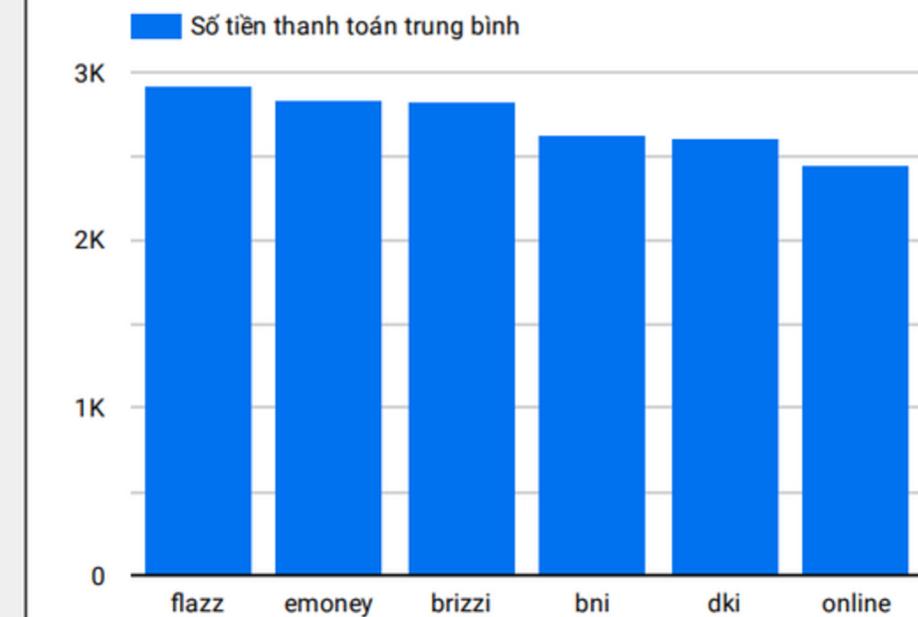
Looker

Thống kê phân tích doanh thu theo ngân hàng.

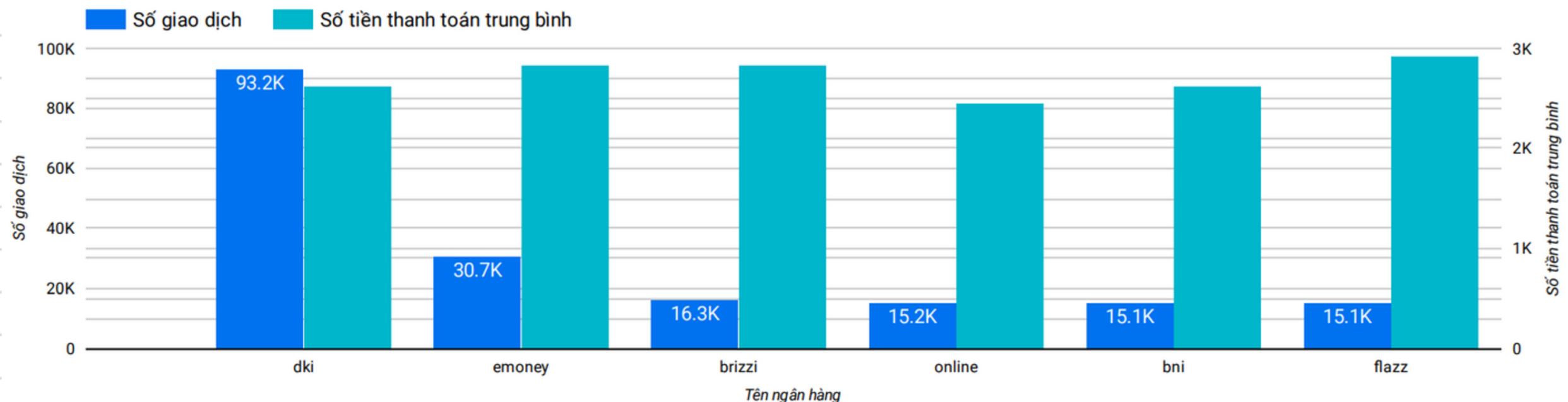
Doanh thu theo ngân hàng



Số tiền thanh toán trung bình theo ngân hàng



Số giao dịch và số tiền thanh toán trung bình trên mỗi giao dịch theo ngân hàng



Data Mining



Thuật toán sử dụng

- Logistic Regression
- Random Forest
- Decision Tree
- Kết quả

| | Model Name | accuracy | roc auc | f1-weighted |
|---|------------------------|-----------------|----------------|--------------------|
| 2 | RandomForestClassifier | 0.981273 | 0.981462 | 0.981280 |
| 1 | DecisionTreeClassifier | 0.977528 | 0.979452 | 0.977564 |
| 0 | LogisticRegression | 0.850187 | 0.848155 | 0.850073 |

Tổng kết