Uber Project

In this project we will extract the data from the source and then upload this data into the sas environment then we will perform the sanity check, clean the data, validate the data by using the different methods and procedures. After cleaning the data we will manipulate our data and create different datasets using the different functions and methods we will create the different variable in those datasets as per our requirement and finally we will combine all those datasets in one dataset and export that dataset perform the visualizations on using the BI tool such as power BI.

- UPLOADING DATA

Uploading data using the infile statement :

We upload data from the flat file such as the text file or csv file into sas with the help of the infile statement. With the infile statement we use infile options such as first obs, dlm, dsd, missover, trunkover to avoid any data loss

CODE:

```
data uber_project(drop= store_and_fwd_flag congestion_surcharge airport_fee) ;

infile "/home/u63650699/sasuser.v94/Uber Project/yellow_tripdata_2017-01.csv"
firstobs=2 dsd;
```

First obs defines that the data should read from the perticular line number

```
input VendorID pickup_date pickup_time dropoff_date dropoff_time passenger_count trip_distance
RatecodeID store_and_fwd_flag PULocationID DOLocationID payment_type fare_amount extra
mta_tax
tip_amount tolls_amount improvement_surcharge total_amount congestion_surcharge $ airport_fee
$;
```

Input is used to define the variables name

```
informat pickup_date date11. pickup_time time11. dropoff_date date11. dropoff_time time11.;
```

Informat is used to read the std data such as date, time amount that contain special character.

```
format pickup_date ddmmyy10. pickup_time time11. dropoff_date ddmmyy10. dropoff_time time11.;
```

Format is used to display the display the data in the form of date, time etc

```
run;
```

- DATA CLEANING AND DATA VALIDATION

Finding missing values:

First we check if there is a missing value in the Vendor ID variable or not, to check this we will use the proc freq, it is one of a way to find a missing values in a variable

CODE:

```
proc freq data= uber_project;
table VendorID;
run;
```

Using this code we see that there are thirteen missing values in the vendor id variable so we will simply delete those observation as it will not significantly impacting our data if we delete those observations. We can delete the observations using the if then statement.

The proc contents provides the details about the Variable position, Variable name, Data Type, Variable Length, format, Informat of the variables and many other informations.

Creating a new dataset and deleting missing values:

CODE:

```
data main;
set uber_project;
if missing(VendorID) then delete;
run;
```

Checking the data without opening it. Using the proc contents we can see our dataset without opening it.

CODE:

```
proc contents data=main; run;
```

Checking if there is any missing value in the vandor id or not in main dataset:

CODE:

```
proc freq data= main;
table VendorID;
run;
```

Sorting the data and removing the duplicates value :

CODE:

```
proc sort data= main nodup; (Nodup removes duplicate records where every field is
duplicate
)
by VendorID;
run;
```

- DATA MANIPULATION

Creating fact table: This is fact table that contain the primary key and other useful variables.

CODE:

```
data Fact;
retain Fact_ID VendorID fare_amount extra mta_tax tip_amount tolls_amount
improvement_surcharge total_amount;
set main(drop=pickup_date pickup_time dropoff_date dropoff_time passenger_count
trip_distance
RatecodeID PULocationID DOLocationID payment_type );
Fact_ID=_N_; (creating a primary key and by using this key we will combine our tables)
run;
```

Creating date and Time Table: Creating a Date and time dimension table which contains the pickup and drop off date time variables:

CODE:

```
data datetime_dim;
retain datetime_ID pickup_date pickup_time dropoff_date dropoff_time Pickup_hour Pickup_day
Pickup_month Pickup_weekday Pickup_Year
dropoff_hour dropoff_day dropoff_month dropoff_weekday dropoff_Year;
set main;
datetime_ID=_N_;
keep datetime_ID pickup_date pickup_time dropoff_date dropoff_time Pickup_hour Pickup_day
Pickup_month Pickup_weekday Pickup_Year
dropoff_hour dropoff_day dropoff_month dropoff_weekday dropoff_Year;
Pickup_hour=hour(pickup_time);
Pickup_day=day(pickup_date);
Pickup_month=month(pickup_date);
Pickup_Year=Year(pickup_date);
Pickup_weekday=weekday(pickup_date);
```

```
dropoff_hour=hour(dropoff_time);
dropoff_day=day(dropoff_date);
dropoff_month=month(dropoff_date);
dropoff_weekday=month(dropoff_date);
dropoff_Year=Year(dropoff_date);
dropoff_weekday=weekday(pickup_date);
run;
```

Creating a Passenger count table: This table show the how many passenger were in the particular vehicle type.

CODE:

```
Data passenger_count_dim;
retain passenger_count_ID passenger_count;
set main;
passenger_count_ID=_N_;
keep passenger_count_ID passenger_count;
Run;
```

Creating trip distance table: This table represents the difference between the pickup location and dropoff location.

CODE:

```
Data trip_distance_dim;
retain trip_distance_ID trip_distance;
set main;
trip_distance_ID=_N_; /*Creating a trip distane id for joing the table in the program*/
keep trip_distance_ID trip_distance;
Run;
```

Creating rate code table: This table represents the different rate code for the different vehicle type or vehicle class.

CODE:

```
data rate_code_dim;
retain rate_code_ID RatecodeID Vehicle_type;
length Vehicle_type $ 20.;
```

```
set main;
rate_code_ID=_N_;
keep rate_code_ID RatecodeID Vehicle_type;
if RatecodeID = 1 then Vehicle_type= "Uber Mini";
else if RatecodeID = 2 then Vehicle_type= "UberX";
else if RatecodeID = 3 then Vehicle_type= "UberXL";
else if RatecodeID = 4 then Vehicle_type= "Uber Black";
else if RatecodeID = 5 then Vehicle_type= "Uber SUV";
else if RatecodeID = 6 then Vehicle_type= "Uber LUX";
else Vehicle_type= "Uber Shared";
Run;
```

Creating payment type table: This table represents the different payment type that a user used to pay for the trip.

CODE:

```
data payment_type_dim;
retain payment_code_ID payment_type payment_type_name;
length payment_type_name $ 40.;
set main;
payment_code_ID=_N_;
Keep payment_code_ID payment_type payment_type_name;
if payment_type = 1 then payment_type_name= "Cash";
else if payment_type = 2 then payment_type_name= "Credit Card";
else if payment_type = 3 then payment_type_name= "Debit Card";
else payment_type_name= "Net Banking";
run;
```

Combining table using proc sql: By using this code we will combine our data and will create a new table for the combined data so that we will export this table and perform the visualizations using the power bi.

CODE:

```
Proc sql;
create table t as
select a.VendorID,
              a.fare_amount,
              a.extra,
              a.mta_tax,
              a.tip_amount,
              a.tolls_amount,
              a.improvement_surcharge,
              a.total_amount,
              b.pickup_date,
              b.pickup_time,
              b.Pickup_hour,
              b.Pickup_day,
              b.Pickup_month,
              b.Pickup_weekday,
              b.Pickup_Year,
              b.dropoff_date,
              b.dropoff_time,
              b.dropoff_hour,
              b.dropoff_day,
              b.dropoff_month,
              b.dropoff_weekday,
              b.dropoff_Year,
              c.passenger_count,
              d.trip_distance,
              e.RatecodeID,
              e.Vehicle_type,
              f.payment_type,
              f.payment_type_name
from Fact as a
left join datetime_dim as b on a.Fact_ID = b.datetime_ID
left join passenger_count_dim as c on a.Fact_ID = c.passenger_count_ID
left join trip_distance_dim as d on a.Fact_ID = d.trip_distance_ID
left join rate_code_dim as e on a.Fact_ID = e.rate_code_ID
left join payment_type_dim as f on a.Fact_ID = f.payment_code_ID
;
quit;
```

Exporting data:

CODE:

```
Proc export data= work.t
outfile="/home/u63650699/sasuser.v94/Uber Project/Final data.xlsx"
dbms=xlsx
replace;
run;
```

Screenshots:

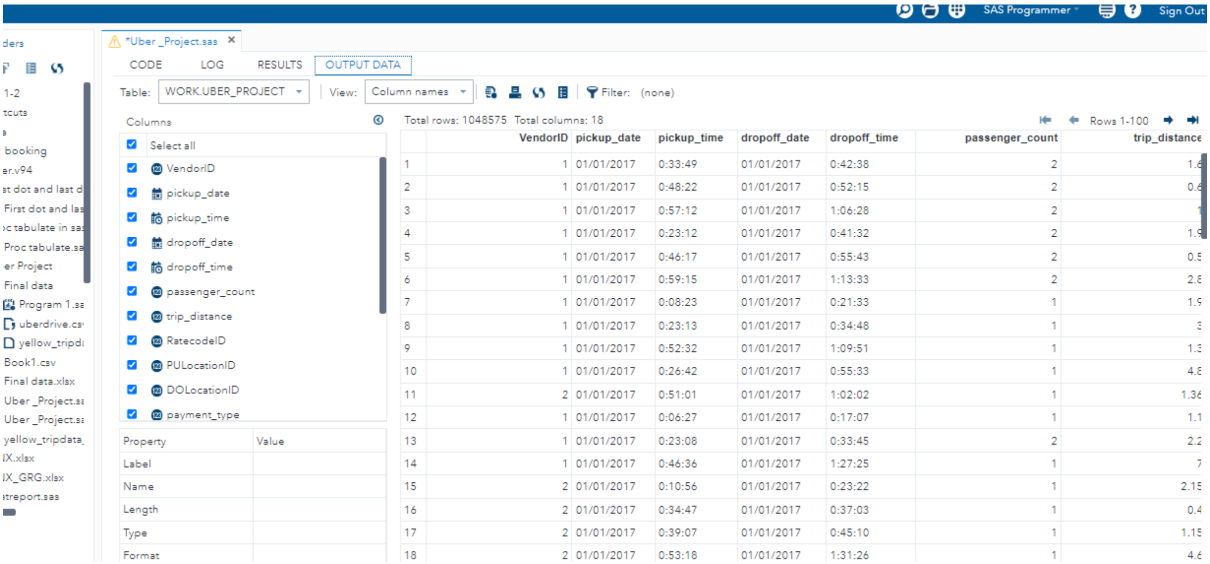


Table: WORK.UBER_PROJECT View: Column names Filter: (none)

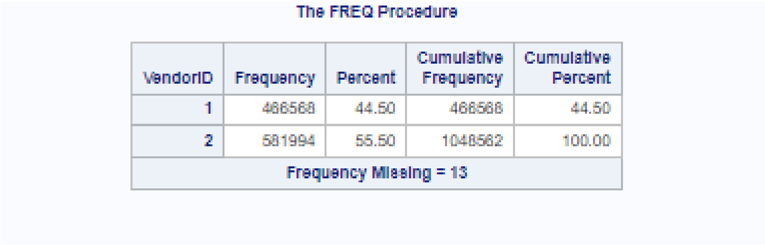
Columns

☑ Select all
☑ VendorID
☑ pickup_date
☑ pickup_time
☑ dropoff_date
☑ dropoff_time
☑ passenger_count
☑ trip_distance
☑ RatecodeID
☑ PULocationID
☑ DOLocationID
☑ payment_type

Property	Value
Label	
Name	
Length	
Type	
Format	

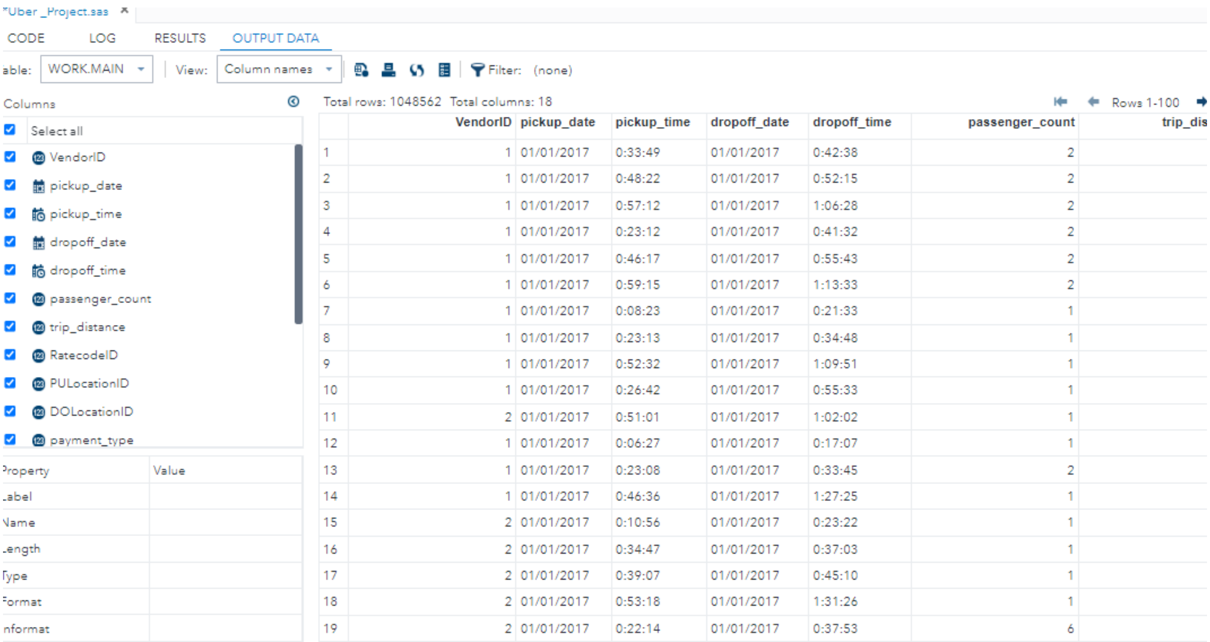
Total rows: 1048575 Total columns: 18 Rows 1-100

	VendorID	pickup_date	pickup_time	dropoff_date	dropoff_time	passenger_count	trip_distance
1	1	01/01/2017	0:33:49	01/01/2017	0:42:38	2	1.6
2	1	01/01/2017	0:48:22	01/01/2017	0:52:15	2	0.6
3	1	01/01/2017	0:57:12	01/01/2017	1:06:28	2	1
4	1	01/01/2017	0:23:12	01/01/2017	0:41:32	2	1.9
5	1	01/01/2017	0:46:17	01/01/2017	0:55:43	2	0.5
6	1	01/01/2017	0:59:15	01/01/2017	1:13:33	2	2.8
7	1	01/01/2017	0:08:23	01/01/2017	0:21:33	1	1.9
8	1	01/01/2017	0:23:13	01/01/2017	0:34:48	1	3
9	1	01/01/2017	0:52:32	01/01/2017	1:09:51	1	1.3
10	1	01/01/2017	0:26:42	01/01/2017	0:55:33	1	4.8
11	2	01/01/2017	0:51:01	01/01/2017	1:02:02	1	1.36
12	1	01/01/2017	0:06:27	01/01/2017	0:17:07	1	1.1
13	1	01/01/2017	0:23:08	01/01/2017	0:33:45	2	2.2
14	1	01/01/2017	0:46:36	01/01/2017	1:27:25	1	7
15	2	01/01/2017	0:10:56	01/01/2017	0:23:22	1	2.15
16	2	01/01/2017	0:34:47	01/01/2017	0:37:03	1	0.4
17	2	01/01/2017	0:39:07	01/01/2017	0:45:10	1	1.15
18	2	01/01/2017	0:53:18	01/01/2017	1:31:26	1	4.6



The FREQ Procedure

VendorID	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	466568	44.50	466568	44.50
2	581994	55.50	1048562	100.00
Frequency Missing = 13				



able: WORK.MAIN View: Column names Filter: (none)

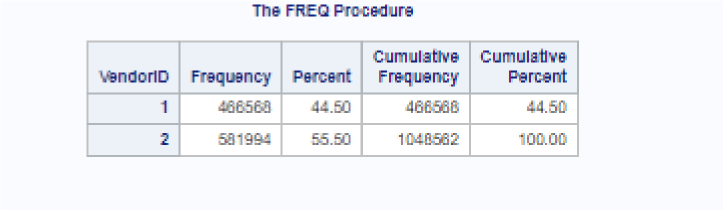
Columns

☑ Select all
☑ VendorID
☑ pickup_date
☑ pickup_time
☑ dropoff_date
☑ dropoff_time
☑ passenger_count
☑ trip_distance
☑ RatecodeID
☑ PULocationID
☑ DOLocationID
☑ payment_type

Property	Value
.abel	
Name	
.ength	
Type	
Format	
nformat	

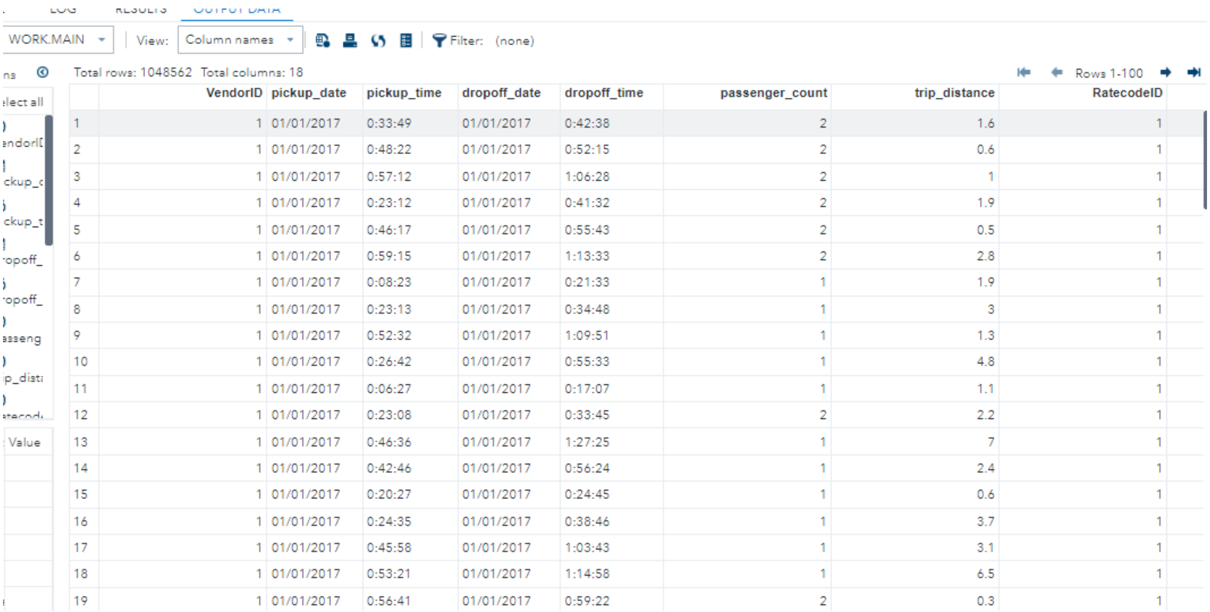
Total rows: 1048562 Total columns: 18 Rows 1-100

	VendorID	pickup_date	pickup_time	dropoff_date	dropoff_time	passenger_count	trip_dis
1	1	01/01/2017	0:33:49	01/01/2017	0:42:38	2	
2	1	01/01/2017	0:48:22	01/01/2017	0:52:15	2	
3	1	01/01/2017	0:57:12	01/01/2017	1:06:28	2	
4	1	01/01/2017	0:23:12	01/01/2017	0:41:32	2	
5	1	01/01/2017	0:46:17	01/01/2017	0:55:43	2	
6	1	01/01/2017	0:59:15	01/01/2017	1:13:33	2	
7	1	01/01/2017	0:08:23	01/01/2017	0:21:33	1	
8	1	01/01/2017	0:23:13	01/01/2017	0:34:48	1	
9	1	01/01/2017	0:52:32	01/01/2017	1:09:51	1	
10	1	01/01/2017	0:26:42	01/01/2017	0:55:33	1	
11	2	01/01/2017	0:51:01	01/01/2017	1:02:02	1	
12	1	01/01/2017	0:06:27	01/01/2017	0:17:07	1	
13	1	01/01/2017	0:23:08	01/01/2017	0:33:45	2	
14	1	01/01/2017	0:46:36	01/01/2017	1:27:25	1	
15	2	01/01/2017	0:10:56	01/01/2017	0:23:22	1	
16	2	01/01/2017	0:34:47	01/01/2017	0:37:03	1	
17	2	01/01/2017	0:39:07	01/01/2017	0:45:10	1	
18	2	01/01/2017	0:53:18	01/01/2017	1:31:26	1	
19	2	01/01/2017	0:22:14	01/01/2017	0:37:53	6	



The FREQ Procedure

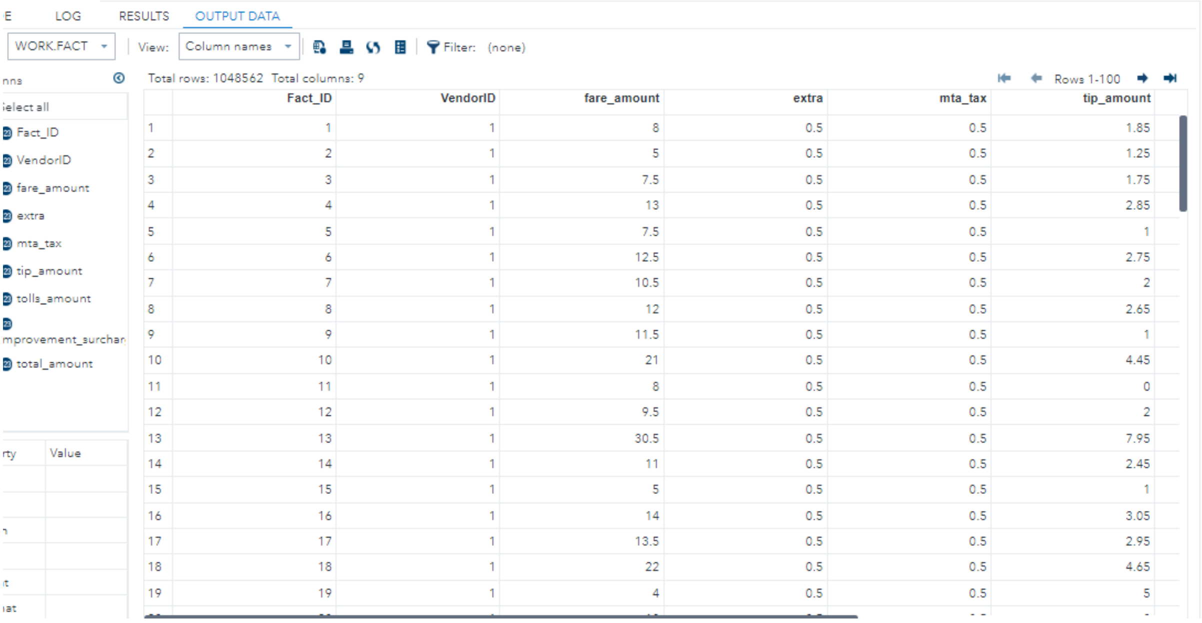
VendorID	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	466568	44.50	466568	44.50
2	581994	55.50	1048562	100.00



WORK.MAIN View: Column names Filter: (none)

Total rows: 1048562 Total columns: 18

	VendorID	pickup_date	pickup_time	dropoff_date	dropoff_time	passenger_count	trip_distance	RatecodeID
1	1	01/01/2017	0:33:49	01/01/2017	0:42:38	2	1.6	1
2	1	01/01/2017	0:48:22	01/01/2017	0:52:15	2	0.6	1
3	1	01/01/2017	0:57:12	01/01/2017	1:06:28	2	1	1
4	1	01/01/2017	0:23:12	01/01/2017	0:41:32	2	1.9	1
5	1	01/01/2017	0:46:17	01/01/2017	0:55:43	2	0.5	1
6	1	01/01/2017	0:59:15	01/01/2017	1:13:33	2	2.8	1
7	1	01/01/2017	0:08:23	01/01/2017	0:21:33	1	1.9	1
8	1	01/01/2017	0:23:13	01/01/2017	0:34:48	1	3	1
9	1	01/01/2017	0:52:32	01/01/2017	1:09:51	1	1.3	1
10	1	01/01/2017	0:26:42	01/01/2017	0:55:33	1	4.8	1
11	1	01/01/2017	0:06:27	01/01/2017	0:17:07	1	1.1	1
12	1	01/01/2017	0:23:08	01/01/2017	0:33:45	2	2.2	1
13	1	01/01/2017	0:46:36	01/01/2017	1:27:25	1	7	1
14	1	01/01/2017	0:42:46	01/01/2017	0:56:24	1	2.4	1
15	1	01/01/2017	0:20:27	01/01/2017	0:24:45	1	0.6	1
16	1	01/01/2017	0:24:35	01/01/2017	0:38:46	1	3.7	1
17	1	01/01/2017	0:45:58	01/01/2017	1:03:43	1	3.1	1
18	1	01/01/2017	0:53:21	01/01/2017	1:14:58	1	6.5	1
19	1	01/01/2017	0:56:41	01/01/2017	0:59:22	2	0.3	1



WORK.FACT View: Column names Filter: (none)

Total rows: 1048562 Total columns: 9

	Fact_ID	VendorID	fare_amount	extra	mta_tax	tip_amount
1	1	1	8	0.5	0.5	1.85
2	2	1	5	0.5	0.5	1.25
3	3	1	7.5	0.5	0.5	1.75
4	4	1	13	0.5	0.5	2.85
5	5	1	7.5	0.5	0.5	1
6	6	1	12.5	0.5	0.5	2.75
7	7	1	10.5	0.5	0.5	2
8	8	1	12	0.5	0.5	2.65
9	9	1	11.5	0.5	0.5	1
10	10	1	21	0.5	0.5	4.45
11	11	1	8	0.5	0.5	0
12	12	1	9.5	0.5	0.5	2
13	13	1	30.5	0.5	0.5	7.95
14	14	1	11	0.5	0.5	2.45
15	15	1	5	0.5	0.5	1
16	16	1	14	0.5	0.5	3.05
17	17	1	13.5	0.5	0.5	2.95
18	18	1	22	0.5	0.5	4.65
19	19	1	4	0.5	0.5	5

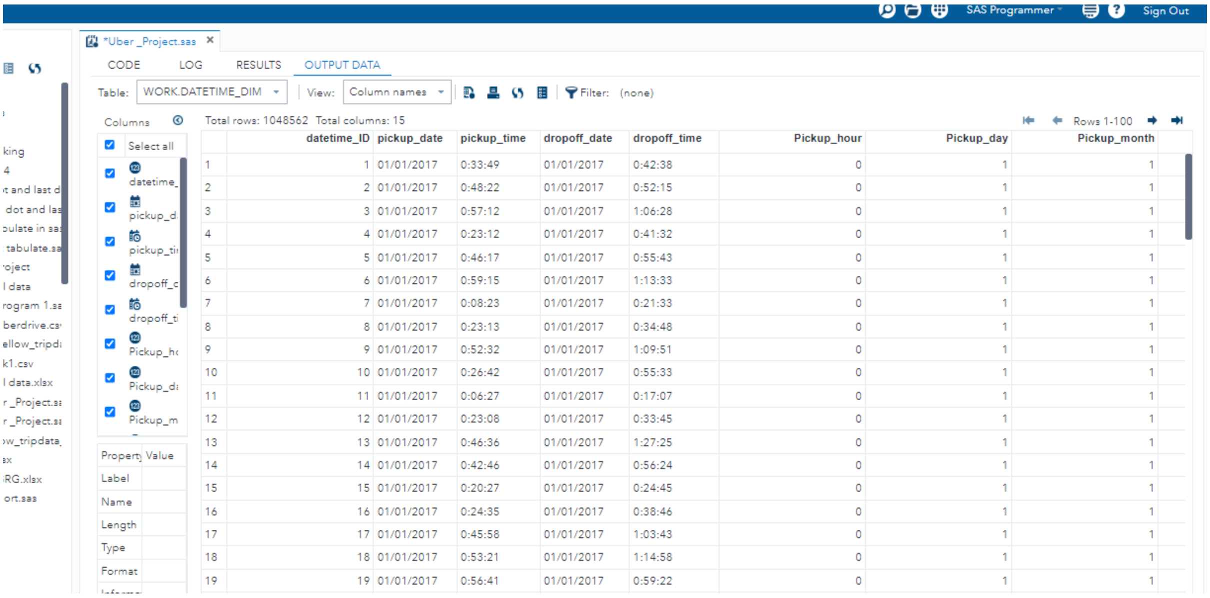


Table: WORK.DATETIME_DIM View: Column names Filter: (none)

Total rows: 1048562 Total columns: 15

Rows 1-100

	datetime_ID	pickup_date	pickup_time	dropoff_date	dropoff_time	Pickup_hour	Pickup_day	Pickup_month
1	1	01/01/2017	0:33:49	01/01/2017	0:42:38	0	1	1
2	2	01/01/2017	0:48:22	01/01/2017	0:52:15	0	1	1
3	3	01/01/2017	0:57:12	01/01/2017	1:06:28	0	1	1
4	4	01/01/2017	0:23:12	01/01/2017	0:41:32	0	1	1
5	5	01/01/2017	0:46:17	01/01/2017	0:55:43	0	1	1
6	6	01/01/2017	0:59:15	01/01/2017	1:13:33	0	1	1
7	7	01/01/2017	0:08:23	01/01/2017	0:21:33	0	1	1
8	8	01/01/2017	0:23:13	01/01/2017	0:34:48	0	1	1
9	9	01/01/2017	0:52:32	01/01/2017	1:09:51	0	1	1
10	10	01/01/2017	0:26:42	01/01/2017	0:55:33	0	1	1
11	11	01/01/2017	0:06:27	01/01/2017	0:17:07	0	1	1
12	12	01/01/2017	0:23:08	01/01/2017	0:33:45	0	1	1
13	13	01/01/2017	0:46:36	01/01/2017	1:27:25	0	1	1
14	14	01/01/2017	0:42:46	01/01/2017	0:56:24	0	1	1
15	15	01/01/2017	0:20:27	01/01/2017	0:24:45	0	1	1
16	16	01/01/2017	0:24:35	01/01/2017	0:38:46	0	1	1
17	17	01/01/2017	0:45:58	01/01/2017	1:03:43	0	1	1
18	18	01/01/2017	0:53:21	01/01/2017	1:14:58	0	1	1
19	19	01/01/2017	0:56:41	01/01/2017	0:59:22	0	1	1

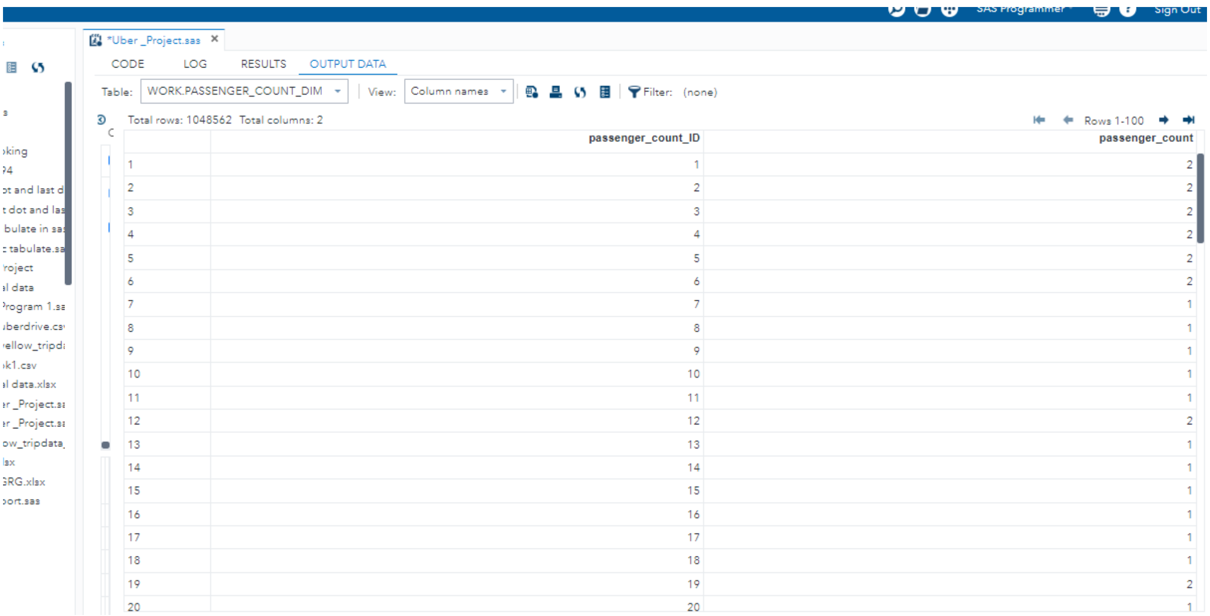


Table: WORK.PASSENGER_COUNT_DIM View: Column names Filter: (none)

Total rows: 1048562 Total columns: 2

Rows 1-100

	passenger_count_ID	passenger_count
1	1	2
2	2	2
3	3	2
4	4	2
5	5	2
6	6	2
7	7	1
8	8	1
9	9	1
10	10	1
11	11	1
12	12	2
13	13	1
14	14	1
15	15	1
16	16	1
17	17	1
18	18	1
19	19	2
20	20	1

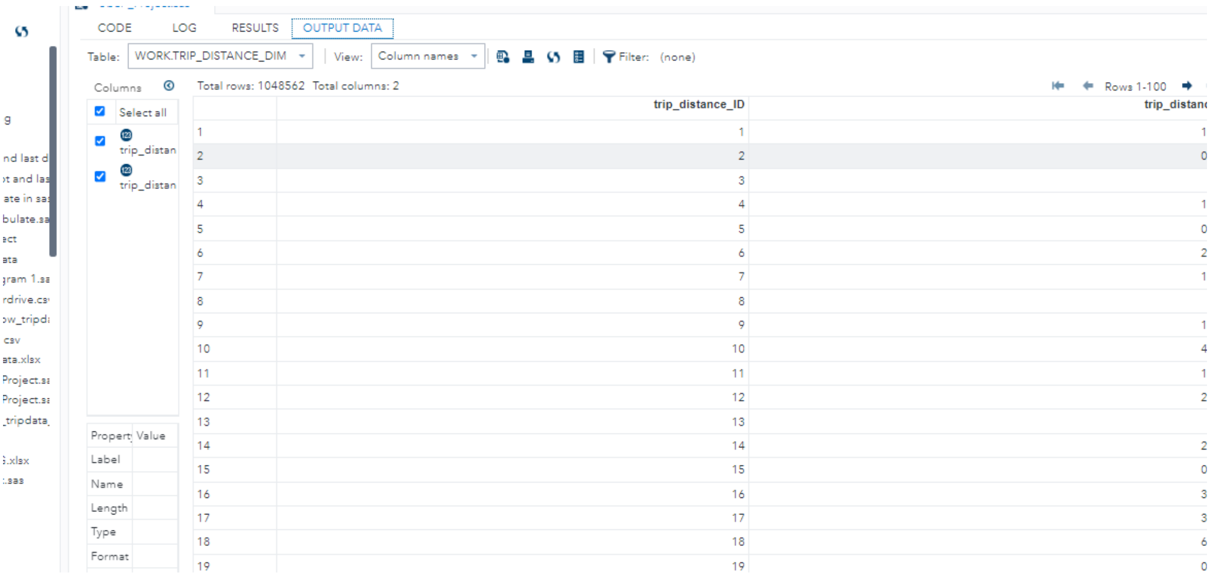


Table: WORK.TRIP_DISTANCE_DIM View: Column names Filter: (none)

Total rows: 1048562 Total columns: 2

	trip_distance_ID	trip_distanc
1	1	1
2	2	0
3	3	
4	4	1
5	5	0
6	6	2
7	7	1
8	8	
9	9	1
10	10	4
11	11	1
12	12	2
13	13	
14	14	2
15	15	0
16	16	3
17	17	3
18	18	6
19	19	0

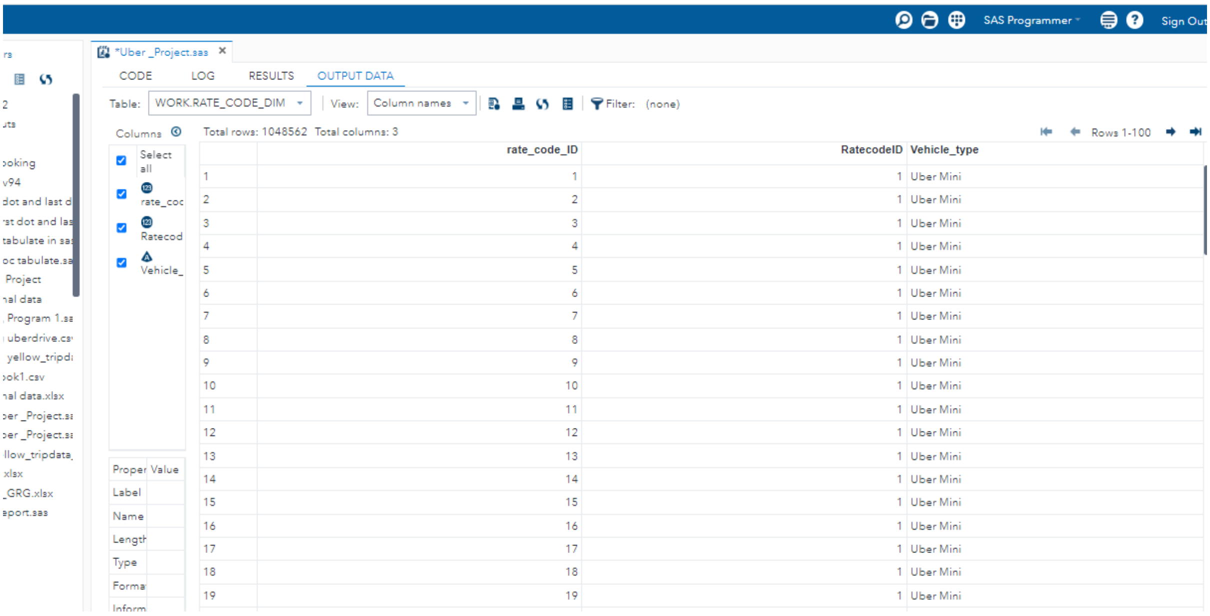
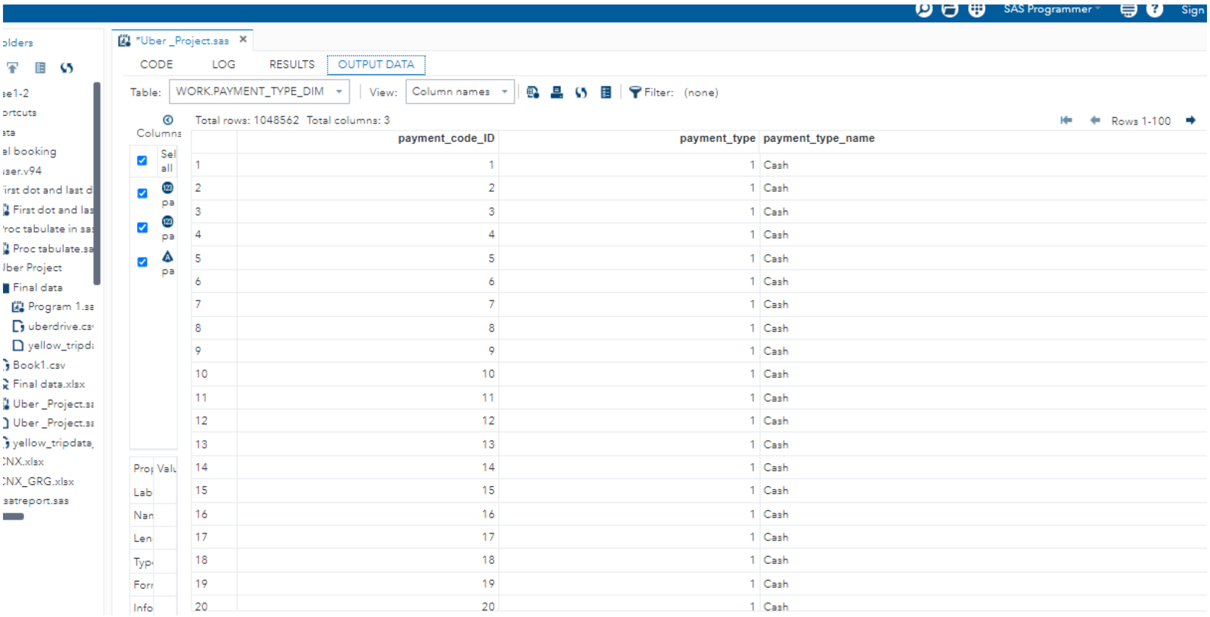


Table: WORK.RATE_CODE_DIM View: Column names Filter: (none)

Total rows: 1048562 Total columns: 3

	rate_code_ID	RatecodeID	Vehicle_type
1	1	1	Uber Mini
2	2	1	Uber Mini
3	3	1	Uber Mini
4	4	1	Uber Mini
5	5	1	Uber Mini
6	6	1	Uber Mini
7	7	1	Uber Mini
8	8	1	Uber Mini
9	9	1	Uber Mini
10	10	1	Uber Mini
11	11	1	Uber Mini
12	12	1	Uber Mini
13	13	1	Uber Mini
14	14	1	Uber Mini
15	15	1	Uber Mini
16	16	1	Uber Mini
17	17	1	Uber Mini
18	18	1	Uber Mini
19	19	1	Uber Mini



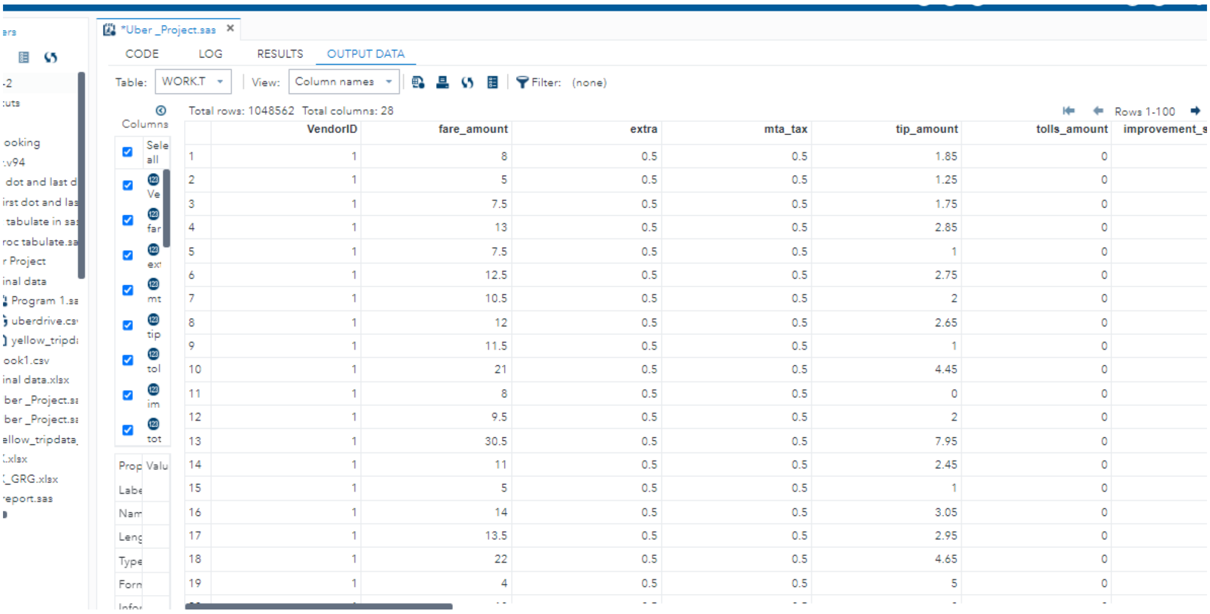
*Uber _Project.sas

CODE LOG RESULTS OUTPUT DATA

Table: WORK.PAYMENT_TYPE_DIM View: Column names Filter: (none)

Total rows: 1048562 Total columns: 3 Rows 1-100

	payment_code_ID	payment_type	payment_type_name
1	1	1	Cash
2	2	1	Cash
3	3	1	Cash
4	4	1	Cash
5	5	1	Cash
6	6	1	Cash
7	7	1	Cash
8	8	1	Cash
9	9	1	Cash
10	10	1	Cash
11	11	1	Cash
12	12	1	Cash
13	13	1	Cash
14	14	1	Cash
15	15	1	Cash
16	16	1	Cash
17	17	1	Cash
18	18	1	Cash
19	19	1	Cash
20	20	1	Cash



*Uber _Project.sas

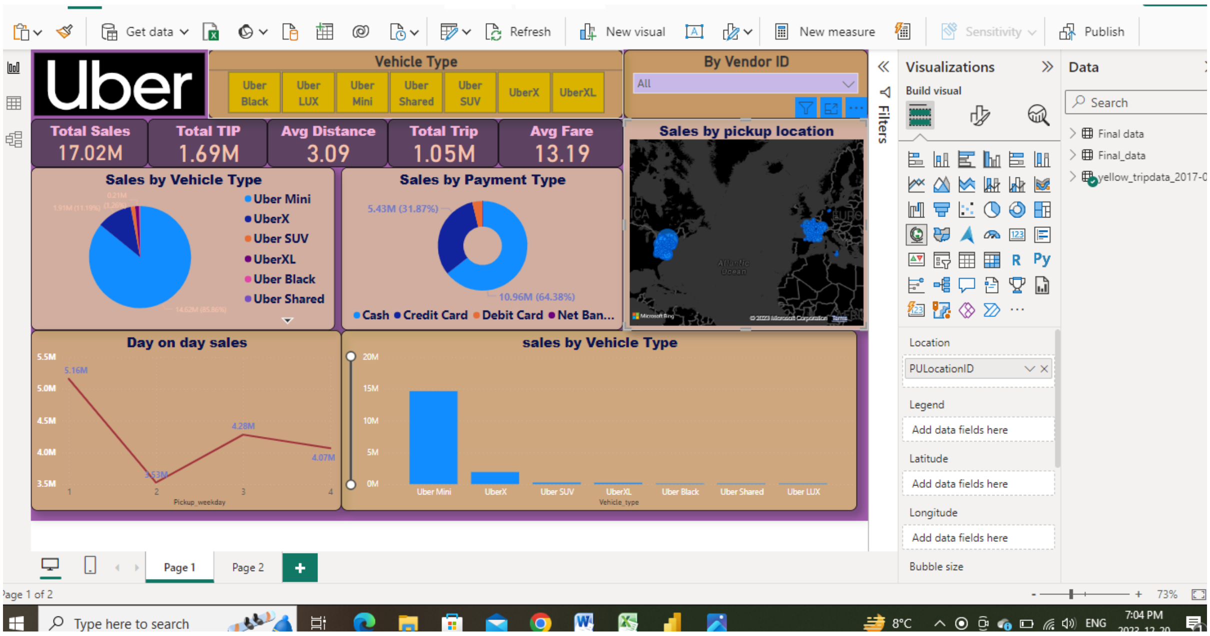
CODE LOG RESULTS OUTPUT DATA

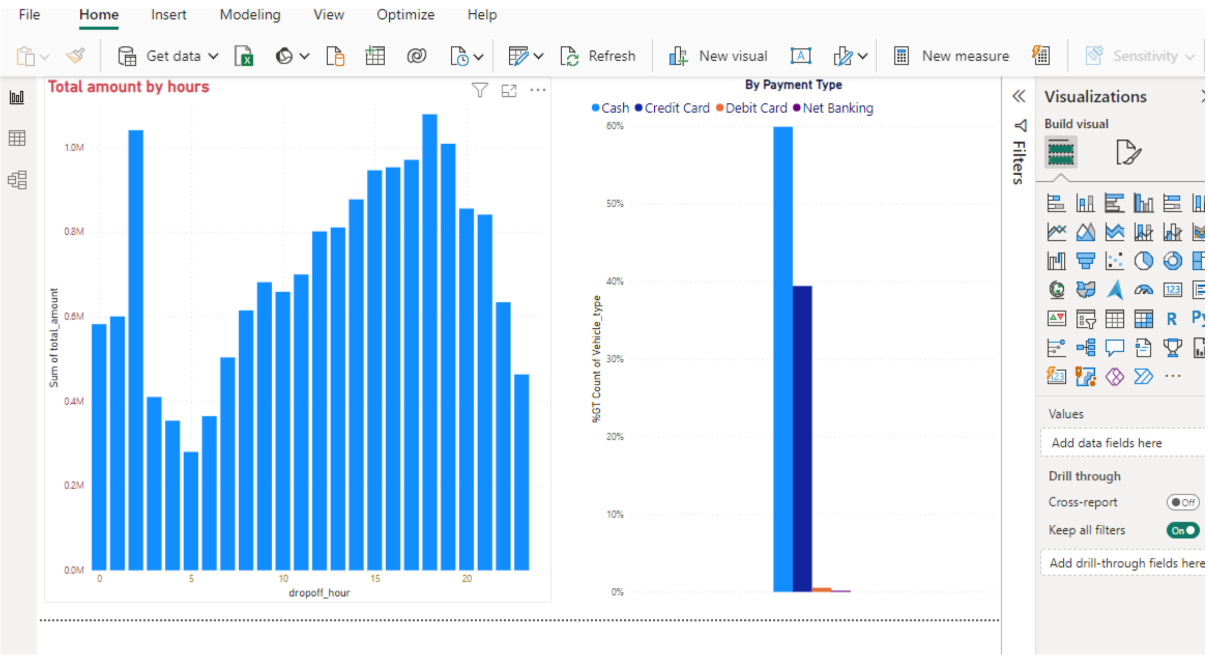
Table: WORK.T View: Column names Filter: (none)

Total rows: 1048562 Total columns: 28 Rows 1-100

	VendorID	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_s
1	1	8	0.5	0.5	1.85	0	
2	1	5	0.5	0.5	1.25	0	
3	1	7.5	0.5	0.5	1.75	0	
4	1	13	0.5	0.5	2.85	0	
5	1	7.5	0.5	0.5	1	0	
6	1	12.5	0.5	0.5	2.75	0	
7	1	10.5	0.5	0.5	2	0	
8	1	12	0.5	0.5	2.65	0	
9	1	11.5	0.5	0.5	1	0	
10	1	21	0.5	0.5	4.45	0	
11	1	8	0.5	0.5	0	0	
12	1	9.5	0.5	0.5	2	0	
13	1	30.5	0.5	0.5	7.95	0	
14	1	11	0.5	0.5	2.45	0	
15	1	5	0.5	0.5	1	0	
16	1	14	0.5	0.5	3.05	0	
17	1	13.5	0.5	0.5	2.95	0	
18	1	22	0.5	0.5	4.65	0	
19	1	4	0.5	0.5	5	0	

Final Output:


Get data
Refresh
New visual
New measure
Sensitivity
Publish
Uber
Vehicle Type
Uber Black
Uber LUX
Uber Mini
Uber Shared
Uber SUV
UberX
UberXL
By Vendor ID
All
Total Sales
17.02M
Total TIP
1.69M
Avg Distance
3.09
Total Trip
1.05M
Avg Fare
13.19
Sales by pickup location
Sales by Vehicle Type
Uber Mini
UberX
Uber SUV
UberXL
Uber Black
Uber Shared
Sales by Payment Type
5.43M (31.87%)
10.96M (64.38%)
Cash
Credit Card
Debit Card
Net Ban...
Day on day sales
5.5M
5.0M
4.5M
4.0M
3.5M
5.16M
4.28M
4.07M
1
2
3
4
Pickup_weekday
sales by Vehicle Type
20M
15M
10M
5M
0M
Uber Mini
UberX
Uber SUV
UberXL
Uber Black
Uber Shared
Uber LUX
Vehicle_type
Visualizations
Build visual
Data
Search
Final data
Final_data
yellow_tripdata_2017-0
Location
PULocationID
Legend
Add data fields here
Latitude
Add data fields here
Longitude
Add data fields here
Bubble size
Filters
Page 1
Page 2
Page 1 of 2
73%
Type here to search
8°C
ENG
7:04 PM


File
Home
Insert
Modeling
View
Optimize
Help
Get data
Refresh
New visual
New measure
Sensitivity
Total amount by hours
1.0M
0.8M
0.6M
0.4M
0.2M
0.0M
Sum of total_amount
0
5
10
15
20
dropoff_hour
By Payment Type
Cash
Credit Card
Debit Card
Net Banking
60%
50%
40%
30%
20%
10%
0%
%GT Count of Vehicle_type
Filters
Visualizations
Build visual
Values
Add data fields here
Drill through
Cross-report
Off
Keep all filters
On
Add drill-through fields here