The report consists of sections in data wrangling, quantitative EDA, and graphical EDA along with comments and storytelling.

All computations are done using the Jupyter Notebook based on Python.
Visualizations are all done using Tableau Public.

Data Wrangling

In addition to the original data’s features, added features of ‘Rides/Requests’ and ‘Online/Total’.

‘Rides/Requests’ records demonstrates the successful rides that actually happen from incoming user requests. The records will tell about customer behavior, for example the patience they have for waiting for their ride to show up.

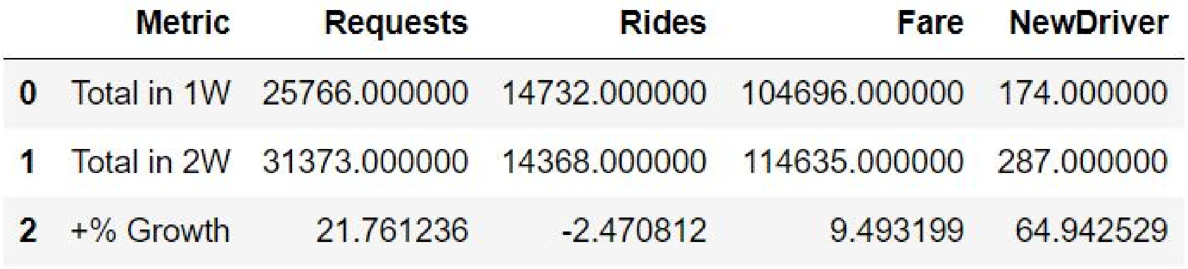
‘Online/Total’ records demonstrates the number of ‘active’ drivers present on the system. This will tell how many registered drivers are (prone to) actually working through Grab.

The csv file for the updated grab data is attached to the email

Quantitative EDA

Growth Analysis

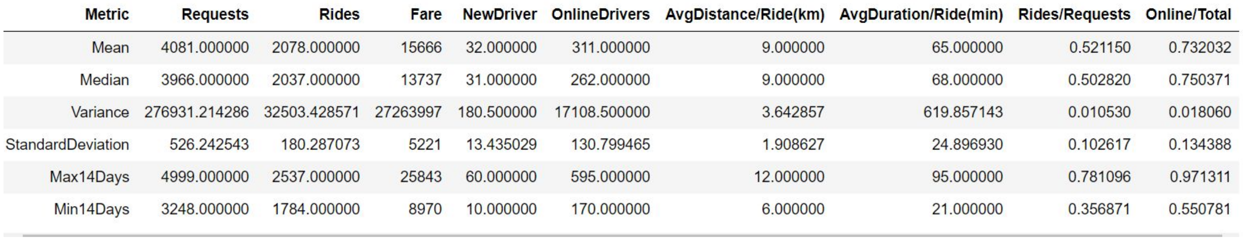
It is essential to point out that the growth of total drivers during the 14 days was up by or plus 210.3%. This value will decrease however if the average number of new drivers do not increase over time.



	Metric	Requests	Rides	Fare	NewDriver
0	Total in 1W	25766.000000	14732.000000	104696.000000	174.000000
1	Total in 2W	31373.000000	14368.000000	114635.000000	287.000000
2	+% Growth	21.761236	-2.470812	9.493199	64.942529

In the above figure, the additional growth from week 1 and week 2 is being computed. Overall, the business is doing an outstanding job because there are many growths. The number of increase in new drivers from week 1 is outstanding with 64% growth. However, the number of rides are almost not changing at all. Even though the requests are going up (which is benefit anyway), if the number of rides does not change then no increase in incoming technically. That’s why there is a not a significant growth in fare.

Computation of the mean, median, variance, standard deviation, maximum and minimum is shown for each feature. *Note than computation for Total Drivers is not present because it is* *futile and has no meaning since Total Drivers are always increasing. CSV file for the computations are attached to the email.*



Metric	Requests	Rides	Fare	NewDriver	OnlineDrivers	AvgDistance/Ride(km)	AvgDuration/Ride(min)	Rides/Requests	Online/Total
Mean	4081.000000	2078.000000	15666	32.000000	311.000000	9.000000	65.000000	0.521150	0.732032
Median	3966.000000	2037.000000	13737	31.000000	262.000000	9.000000	68.000000	0.502820	0.750371
Variance	276931.214286	32503.428571	27263997	180.500000	17108.500000	3.642857	619.857143	0.010530	0.018060
StandardDeviation	526.242543	180.287073	5221	13.435029	130.799465	1.908627	24.896930	0.102617	0.134388
Max14Days	4999.000000	2537.000000	25843	60.000000	595.000000	12.000000	95.000000	0.781096	0.971311
Min14Days	3248.000000	1784.000000	8970	10.000000	170.000000	6.000000	21.000000	0.356871	0.550781

Mean and Median: Generally, the mean and median in the data will be different because mean is the expected value and median is the value that separates the higher half from the lower half. If the mean and median are equal or somewhat equal, then the distribution is symmetrical. I will demonstrate with a box and whisker plot later.

Comment: Grab needs to solve on the low ratio between rides and requests. With a mean of success rides around 52%, there is room for improvement and solutions. What Grab cannot solve is the traffic on the road. Users will get impatient due to the long waiting time, and will start cancelling. For example, Grab may choose to focus on giving compensations, such as giving them some discount coupon to use for further rides under the term that users have to wait more than the app's predicted time

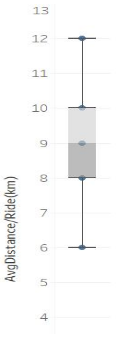
Variance and Standard Deviation: The variance is in squared units so we will be looking at standard deviation. It tells you the average swing from the mean. Generally, if the the percentage of swing over the mean is more than 50% that means the data is not stable and volatile. Features that contain volatile data (around 30% swing) are 'Fare','New Drivers','Online Drivers', and 'AvgDuration/Ride(min)'.

Maximum and Minimum: shows maximum and minimum of records in 14 days.

Box and Whisker Plots

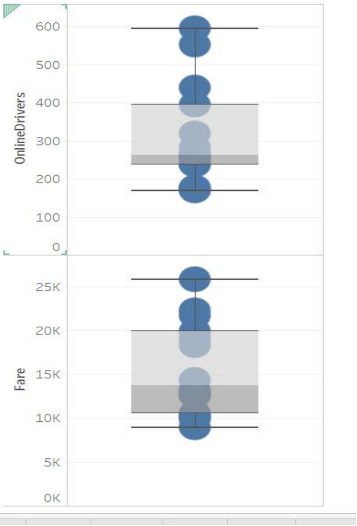
- The middle is the median and corresponds to 50th percentile
- The box's boundary in the middle governing the shaded area, corresponds to 75th and 25th percentile
- If the distance between the point and 75th percentile is more than or equal to 2 IQR, then it is an outlier. This is an estimate though, because there is no official meaning to outlier. 2 IQR is a common criteria. Anything outside the whisker are considered generally to be outliers.

(i) 'AvgDistance/Ride(km)'



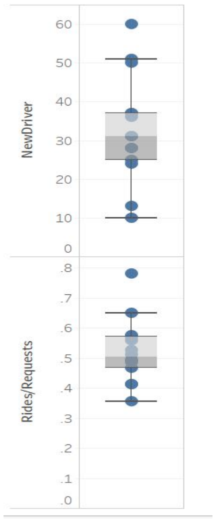
This is the distribution of ‘AvgDistance/Ride(km)’. The mean and median of the feature records are equal, that’s why its distribution is symmetrical.

(ii) ‘Fare’ and ‘OnlineDrivers’



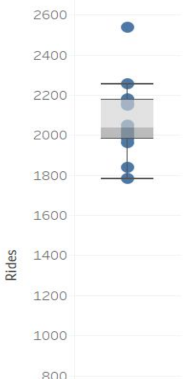
This is the distribution of ‘Fare’ and ‘OnlineDrivers’. The mean and median for these are significantly different and you can see the distribution is skewed. This is fair because in the real world, the amount of fare and online drivers are likely to be fluctuating alot anyways.

(iii) Potential Outliers: ‘Rides/Requests’ and ‘NewDriver’



In the plots, there are potential outliers. They are classified as potential because they are outliers that are around 2 IQR from the 75th percentile. These outliers, however, are not considered to be erroneous because they are not clearly abnormal values.

(iv) ‘Rides’



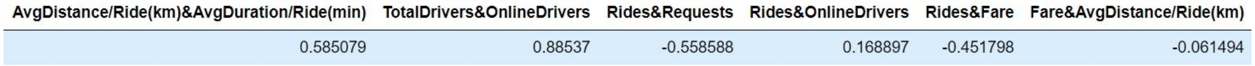
This is an outlier that has exceeded 2 IQR. The outlier is considered to be erroneous because it is way out of range and can be neglected in further analysis.

Pearson Correlation Coefficients

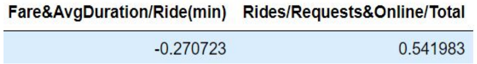
The computation of the coefficients will tell you how correlated certain features are to one another. High positive correlation means that one feature seems to follow one another. The converse means that if one feature tends to decrease the other increases, and vice-versa as well.

Not every combination of features are taken into computation of the coefficient; only the ones that are considered to have impact on the business.

Below are the coefficients corresponding to the combinations.



AvgDistance/Ride(km)&AvgDuration/Ride(min)	TotalDrivers&OnlineDrivers	Rides&Requests	Rides&OnlineDrivers	Rides&Fare	Fare&AvgDistance/Ride(km)
0.585079	0.88537	-0.558588	0.168897	-0.451798	-0.061494

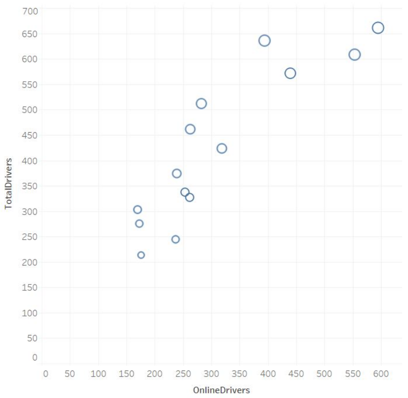


Fare&AvgDuration/Ride(min)	Rides/Requests&Online/Total
-0.270723	0.541983

Notable combinations are average distance per ride & average duration per ride as well as rides/requests & online/total drivers. This is sensible because duration increases as distance increases and the number of successful rides depend on the number of online drivers.

With a coefficient of 0.885, this means that online drivers increases as total drivers increase. This is reassuring for the business because they are trying to increase total drivers so that the number of drivers available for passengers gain more liquidity. Below is a scatter plot of the combination.

There are some expected results. Intuitively, Rides and Request have to have a correlation as well as Rides and Fare. This uncorrelation can be due to the difference in performance for weekday and weekend.



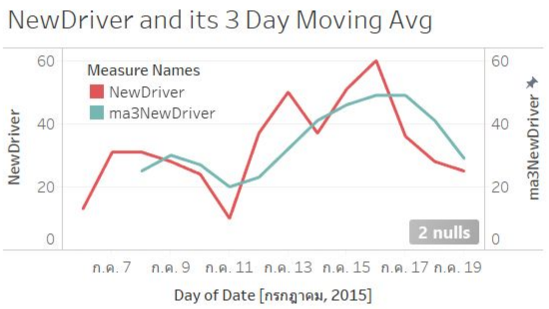
Graphical EDA

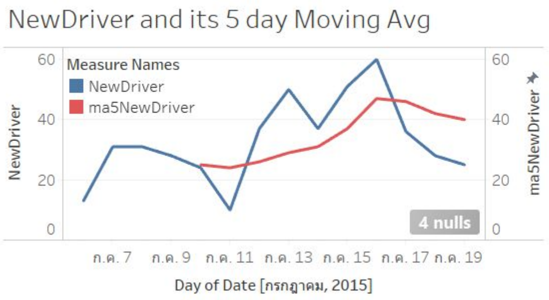
As a better alternative than representing the data with a histogram, illustrating the data (feature records) trend along side its moving averages (MA) has more opportunity for insights.

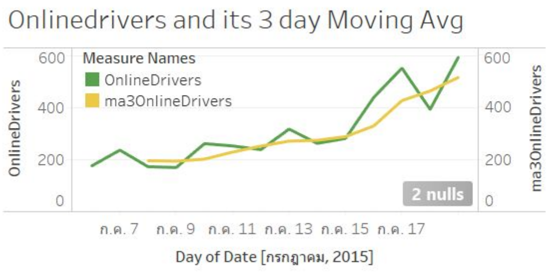
Note that moving averages for total drivers is not included because total drivers would be forever increasing.

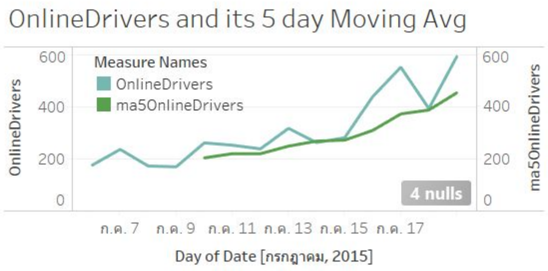
In a range of 14 days, the moving averages for 3 days and 5 days will be conducted and compared.

Examples of comparisons









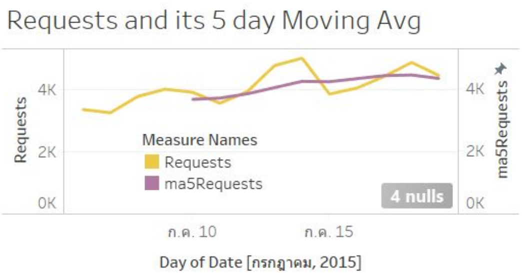
The graphs of features ‘OnlineDrivers’ and ‘NewDriver’ are illustrated above. The purpose here is to differentiate between 5 day and 3 day MA. The graphs depict that the 5 day MA will be smoother than 3 day MA because it covers for wider range of values and the averages are more unlikely to fluctuate.

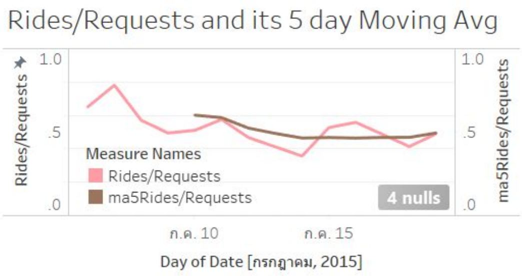
The csv file for the 5 day moving averages is attached to the email

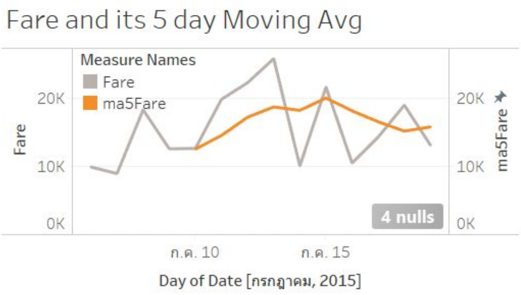
What does moving averages say

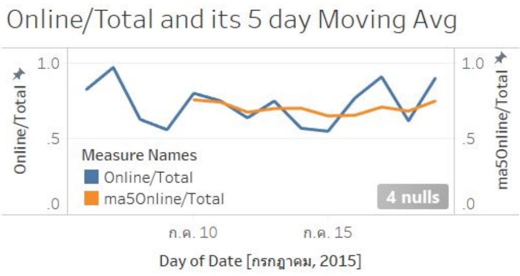
If the original data trend line is above the trend line of the moving averages, then the business is doing its job in increasing growth. This simply means that the business is doing an above average job. Furthermore, even though the data is fluctuating, if overtime the moving averages are increasing, then it is growth.

A good example of where the business is doing an above average job is acquisition of online drivers. In the online drivers graph above, the original trend is sitting above the moving averages at all times and plus the MA trend is increasing. Again, a good job for liquidating online drivers. For performance of acquiring new drivers, it is growth until mid-second week, before dipping down, causing the MA trend to decrease.









MA5 graphs for the rest of the features are illustrated above

Discussion:

Fare: Incoming fare has high fluctuation, so analyzing its moving averages is very important so that one can tell if the growth (while fluctuating) is present or not. It is unclear of what direction will the fare go.

Requests: Moving averages show that there is growth in requests and although it seems like trivial growth, the scale of the graph is 2000 units for each line. However, the MA trend has reach a smooth peak and is starting to slow down its growth.

Rides/Requests: This value is decreasing in week 1 but begins to smooth out at mid second week, which as of now looks like there is a reversal in trend signaling for growth.

Online/Total: This value is quite stable during the two weeks. Referring to mean of 0.73 with 0.13 standard deviation, which makes sense why it is quite stable.