* What's the average speed of yellow cabs on odd days of the week (Monday, Wednesday, Friday, Sunday)?

**Description – Here, I realized that it’s better to represent average speed against WeekDay and therefore I grouped by WeekDays and following is the query. Besides, I realized data cleaning is utterly important in such huge data and therefore the filter conditions such as on trip\_distance and dropoff\_distance are found out after analyzing the data. I hope I have used the right filter conditions.**

**Query –**

select

Dayofweek(pickup\_datetime) as day\_of\_week,

sum(trip\_distance)/(sum((dropoff\_datetime)-(pickup\_datetime))/3600000000) avg\_speed

FROM [nyc-tlc:yellow.trips]

where dropoff\_datetime>pickup\_datetime

and trip\_distance>0

and (Dayofweek(pickup\_datetime) in (1,2,4,6) and Dayofweek(dropoff\_datetime) in (1,2,4,6))

group by 1

order by 1

**RESULT –**

| **Row** | **day\_of\_week** | **avg\_speed** |
| --- | --- | --- |
| 1 | 1 | 28.905435829646645 |
| 2 | 2 | 28.82722614004543 |
| 3 | 4 | 23.418126721425125 |
| 4 | 6 | 28.809875474750946 |

* Using data from the first half of 2015, compute the difference in average fare-per-mile for yellow cabs vs. green cabs. In the month of March 2015, what was the average fare-per-mile for green cabs by day of week? Your results should include a date, yellow fare, and green fare.

**Description – Here, I felt it’s important to crunch data in the months and further into weeks for analyzing the difference in average fare-per-mile for yellow cabs vs. green cabs. Therefore, I grouped on the month and week to analyze it in a better way. To cleanup the data, I used the filter conditions ( trip\_distance>0 and fare\_amount>0). Following is the query with respect to the following question. Moreover, at the last of the document, I have mentioned the query which will reflect the consolidated data.**

**QUERY of First Section** – Using data from the first half of 2015, , Compute the difference in average fare-per-mile for yellow cabs vs. green cabs.

SELECT

t1.yellow\_month as Month,

t1.yellow\_week as Day\_Of\_Week,

t1.avg\_yellow\_fare as yellow\_fare,

t2.avg\_green\_fare as green\_fare,

abs(t2.avg\_green\_fare-t1.avg\_yellow\_fare) as yello\_green\_fair\_diff

FROM

(SELECT

month(pickup\_datetime) as yellow\_month,Dayofweek(pickup\_datetime) as yellow\_week,sum(fare\_amount)/sum(trip\_distance) as avg\_yellow\_fare

FROM [nyc-tlc:yellow.trips]

WHERE month(pickup\_datetime)<=6 and year(pickup\_datetime)=2015

and year(dropoff\_datetime)=2015 and dropoff\_datetime>pickup\_datetime

and trip\_distance>0 and fare\_amount>0

group by 1,2

order by 1,2

) t1

inner JOIN

(SELECT

month(pickup\_datetime) as green\_month,Dayofweek(pickup\_datetime) as green\_week,sum(fare\_amount)/sum(trip\_distance) as avg\_green\_fare

FROM [nyc-tlc:green.trips\_2015]

WHERE month(pickup\_datetime)<=6 and year(dropoff\_datetime)=2015

and year(pickup\_datetime)=2015 and dropoff\_datetime>pickup\_datetime

and trip\_distance>0 and fare\_amount>0

group by 1,2

order by 1,2

) t2 on

t1.yellow\_month=t2.green\_month and t1.yellow\_week=t2.green\_week

order by 1,2

**RESULT** – Compute the difference in average fare-per-mile for yellow cabs vs. green cabs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Month | Day\_Of\_Week | yellow\_fare | green\_fare | yello\_green\_fair\_diff |
| 1 | 1 | 0.464136513 | 3.986385472 | 3.52224896 |
| 1 | 2 | 3.968635818 | 4.217589921 | 0.248954103 |
| 1 | 3 | 4.326810853 | 4.303006751 | 0.023804102 |
| 1 | 4 | 2.352113569 | 4.303931953 | 1.951818385 |
| 1 | 5 | 1.805504314 | 4.195574985 | 2.390070671 |
| 1 | 6 | 0.339696866 | 4.291056782 | 3.951359916 |
| 1 | 7 | 1.240769599 | 4.111343714 | 2.870574115 |
| 2 | 1 | 0.460809637 | 4.016632546 | 3.555822909 |
| 2 | 2 | 0.442236206 | 4.286947645 | 3.844711439 |
| 2 | 3 | 0.773927699 | 4.416746594 | 3.642818895 |
| 2 | 4 | 1.756093347 | 4.421396864 | 2.665303517 |
| 2 | 5 | 1.085824352 | 4.366983164 | 3.281158812 |
| 2 | 6 | 0.382247871 | 4.422911128 | 4.040663257 |
| 2 | 7 | 4.361292216 | 4.201787147 | 0.159505069 |
| 3 | 1 | 0.878930909 | 4.070536591 | 3.191605681 |
| 3 | 2 | 0.495509175 | 4.280137427 | 3.784628252 |
| 3 | 3 | 0.603085435 | 4.431770592 | 3.828685157 |
| 3 | 4 | 0.411383418 | 4.342500415 | 3.931116997 |
| 3 | 5 | 1.567218772 | 4.338788207 | 2.771569436 |
| 3 | 6 | 1.135649496 | 4.368534134 | 3.232884638 |
| 3 | 7 | 0.963327748 | 4.150144921 | 3.186817173 |
| 4 | 1 | 1.407850461 | 4.019731669 | 2.611881208 |
| 4 | 2 | 0.255416529 | 4.220366755 | 3.964950226 |
| 4 | 3 | 0.717344372 | 4.286624785 | 3.569280412 |
| 4 | 4 | 0.2398321 | 4.31140348 | 4.071571381 |
| 4 | 5 | 0.348452315 | 4.312311888 | 3.963859573 |
| 4 | 6 | 0.326938807 | 4.245181709 | 3.918242902 |
| 4 | 7 | 1.083164047 | 4.167775188 | 3.084611142 |
| 5 | 1 | 0.524774428 | 4.035975772 | 3.511201344 |
| 5 | 2 | 3.618811082 | 4.142356441 | 0.523545359 |
| 5 | 3 | 1.164346706 | 4.30255411 | 3.138207404 |
| 5 | 4 | 0.806283399 | 4.314763395 | 3.508479996 |
| 5 | 5 | 3.511818103 | 4.341019893 | 0.82920179 |
| 5 | 6 | 2.991932978 | 4.326130581 | 1.334197603 |
| 5 | 7 | 0.475687527 | 4.155313363 | 3.679625836 |
| 6 | 1 | 1.424659323 | 4.058195658 | 2.633536335 |
| 6 | 2 | 3.00323163 | 4.249178501 | 1.245946872 |
| 6 | 3 | 0.86871238 | 4.348158413 | 3.479446033 |
| 6 | 4 | 1.972431702 | 4.35397294 | 2.381541238 |
| 6 | 5 | 1.502478389 | 4.326187132 | 2.823708743 |
| 6 | 6 | 0.494561177 | 4.342561919 | 3.848000742 |
| 6 | 7 | 1.10414513 | 4.20215451 | 3.09800938 |

**QUERY of Next Section** - In the month of March 2015, what was the average fare-per-mile for green cabs by day of week?

I have just added additional condition(Month=3) to represent green cab data for March month.

select t3.Day\_Of\_Week as Day\_Of\_Week,t3.green\_fare as green\_fare

from (

SELECT

t1.yellow\_month as Month,

t1.yellow\_week as Day\_Of\_Week,

t1.avg\_yellow\_fare as yellow\_fare,

t2.avg\_green\_fare as green\_fare,

abs(t2.avg\_green\_fare-t1.avg\_yellow\_fare) as yello\_green\_fair\_diff

FROM

(SELECT

month(pickup\_datetime) as yellow\_month,Dayofweek(pickup\_datetime) as yellow\_week,sum(fare\_amount)/sum(trip\_distance) as avg\_yellow\_fare

FROM [nyc-tlc:yellow.trips]

WHERE month(pickup\_datetime)<=6 and year(pickup\_datetime)=2015

and year(dropoff\_datetime)=2015 and dropoff\_datetime>pickup\_datetime

and trip\_distance>0 and fare\_amount>0

group by 1,2

order by 1,2

) t1

inner JOIN

(SELECT

month(pickup\_datetime) as green\_month,Dayofweek(pickup\_datetime) as green\_week,sum(fare\_amount)/sum(trip\_distance) as avg\_green\_fare

FROM [nyc-tlc:green.trips\_2015]

WHERE month(pickup\_datetime)<=6 and year(dropoff\_datetime)=2015

and year(pickup\_datetime)=2015 and dropoff\_datetime>pickup\_datetime

and trip\_distance>0 and fare\_amount>0

group by 1,2

order by 1,2

) t2 on

t1.yellow\_month=t2.green\_month and t1.yellow\_week=t2.green\_week

order by 1,2

) as t3 where t3.month=3

RESULTS - In the month of March 2015, what was the average fare-per-mile for green cabs by day of week?

| **Row** | **Day\_Of\_Week** | **green\_fare** |
| --- | --- | --- |
| 1 | 1 | 4.070536590653378 |
| 2 | 2 | 4.2801374267314305 |
| 3 | 3 | 4.431770592124953 |
| 4 | 4 | 4.34250041513264 |
| 5 | 5 | 4.338788207060538 |
| 6 | 6 | 4.368534134417498 |
| 7 | 7 | 4.150144920762452 |

**Important Note**: In second question, I was not clear which date needs to be reflected against every row and therefore I didn’t mess up my data and preferred not to show the date. Moreover, I have written a single query for the second part. However, to fetch the data of “In the month of March 2015, what was the average fare-per-mile for green cabs by day of week”, I have added additional condition (MONTH=3).

I believe the questions are open ended and I analyzed it by breaking the data in months and weeks. But if for second part I’m required to represent the consolidated data then following is the query

Question -

“

Using data from the first half of 2015, , Compute the difference in average fare-per-mile for yellow cabs vs. green cabs.”

**Query** -

SELECT

t1.avg\_yellow\_fare as yellow\_fare,t2.avg\_green\_fare as green\_fare,

abs(t2.avg\_green\_fare-t1.avg\_yellow\_fare) as yello\_green\_fair\_diff

FROM

(SELECT sum(fare\_amount)/sum(trip\_distance) as avg\_yellow\_fare

FROM [nyc-tlc:yellow.trips]

WHERE month(pickup\_datetime)<=6 and year(pickup\_datetime)=2015

and year(dropoff\_datetime)=2015 and dropoff\_datetime>pickup\_datetime

and trip\_distance>0 and fare\_amount>0

) t1

CROSS JOIN

(SELECT sum(fare\_amount)/sum(trip\_distance) as avg\_green\_fare

FROM [nyc-tlc:green.trips\_2015]

WHERE month(pickup\_datetime)<=6 and year(dropoff\_datetime)=2015 and year(pickup\_datetime)=2015

and dropoff\_datetime>pickup\_datetime

and trip\_distance>0 and fare\_amount>0

) t2

**RESULT** –

|  |
| --- |
|  |
| **Row** | **yellow\_fare** | **green\_fare** | **yello\_green\_fair\_diff** |
| 1 | 0.7116357530087779 | 4.239586872752019 | 3.5279511197432414 |