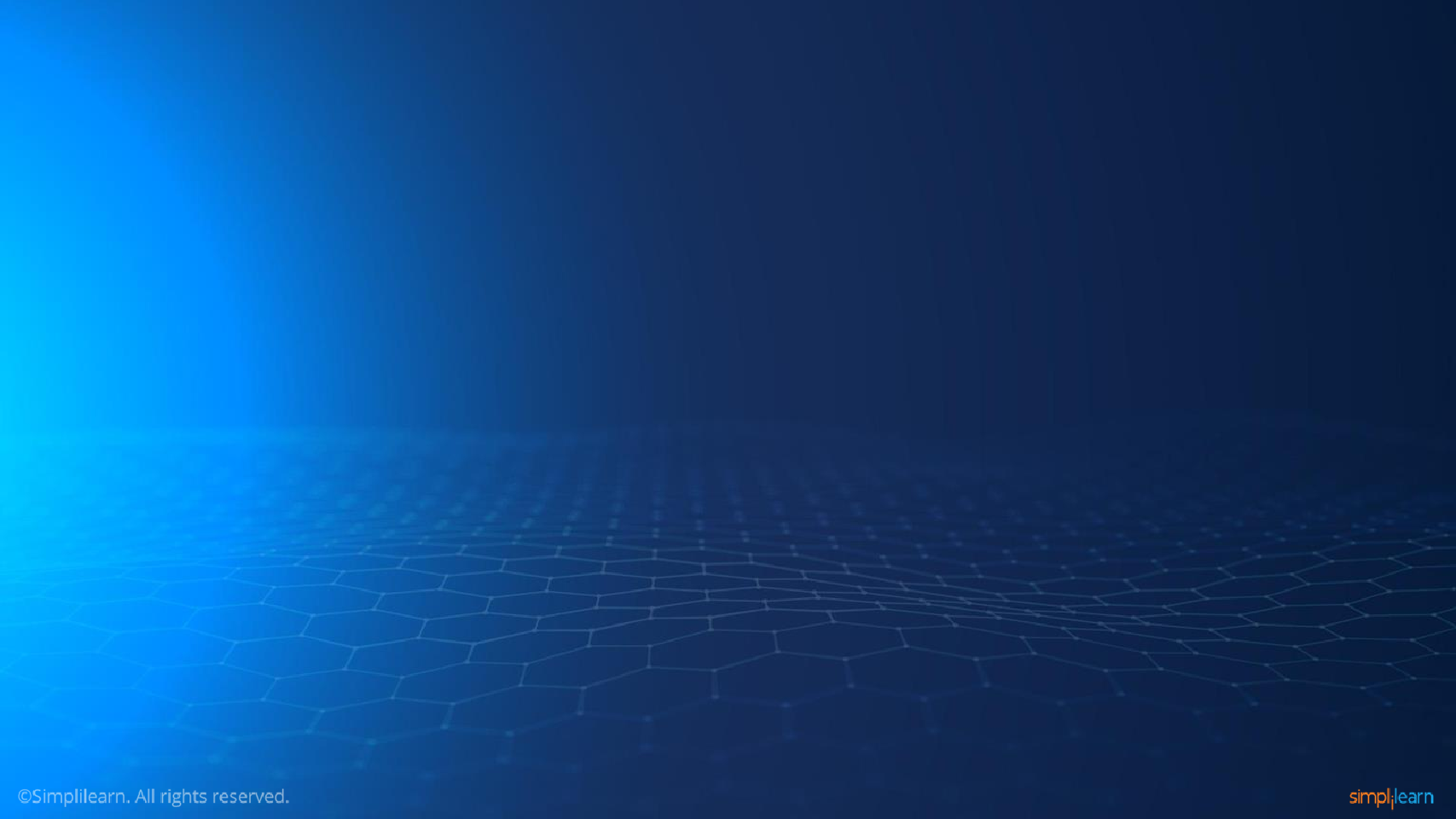


**DataScienceCapstone**



**UnitedStatesAirlinesAnalysis**

# BusinessScenario



Problemstatement:

Accordingtoairtravelconsumerreports,alargeproportionofconsumercomplaintsareaboutfrequentflightdelays.

Outofallthecomplaintsreceivedfromconsumersaboutairlineservices,32%wererelatedto

cancellations,delays,orotherdeviationsfromtheairlines’schedules.

There are unavoidable delays that can be caused by air traffic, no passengers at the airport,weather conditions, mechanical issues, passengers coming from delayed connecting flights,securityclearance,andaircraftpreparation.

Objective:

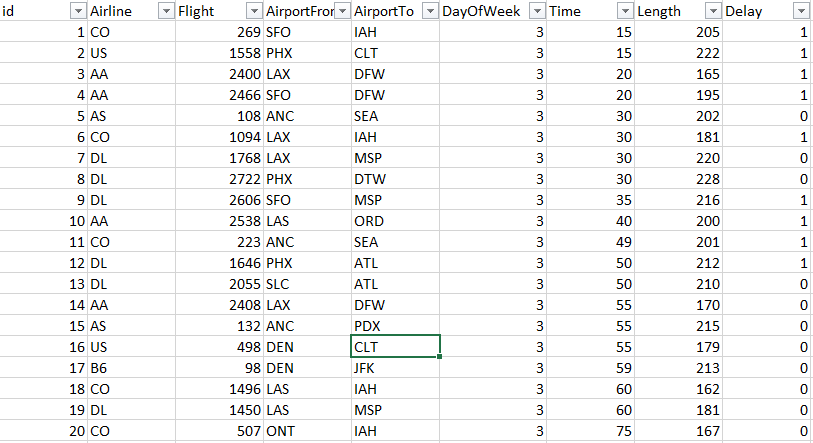
Theobjectiveofthisprojectistoidentifythefactorsthatcontributetoavoidableflightdelays.

Youarealsorequiredtobuildamodeltopredictiftheflightwillbedelayed.

# DatasetSnapshot



**Airlines.xlsx**



# DatasetDescription



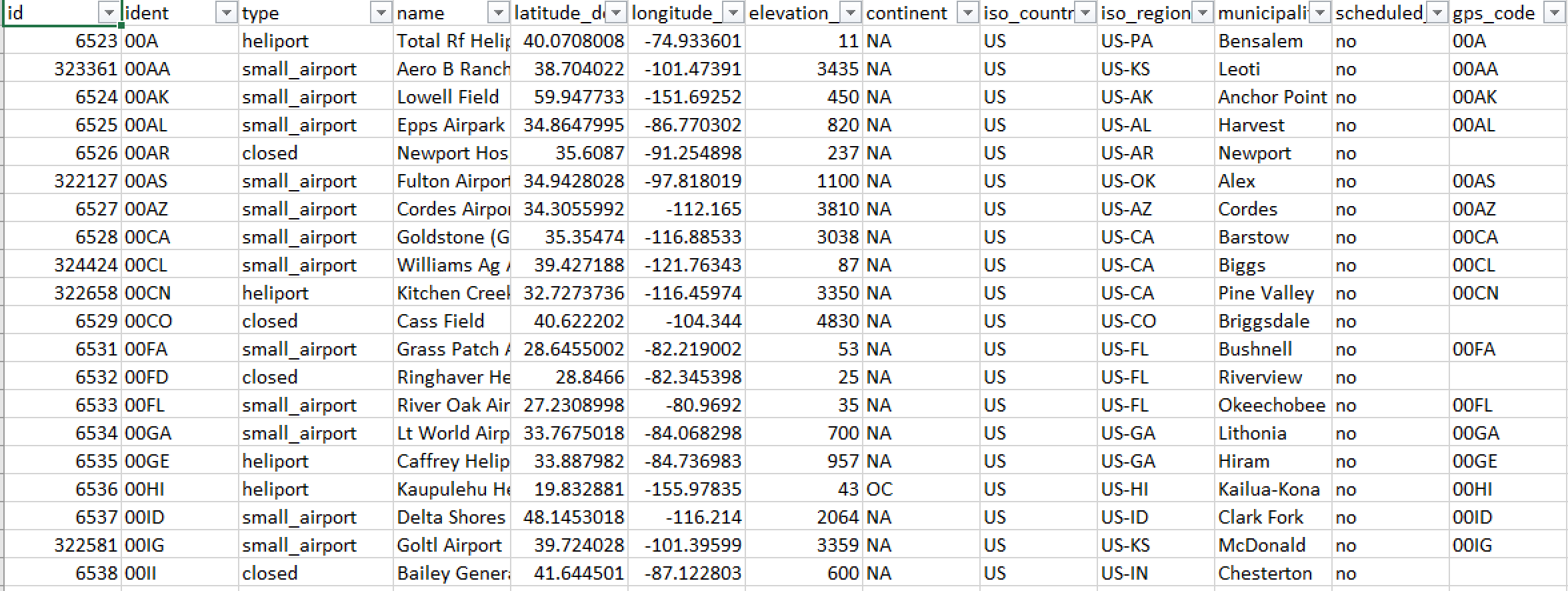
**Airlines.xlsx**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| id | Flightnumber |
| Airline | Typeof commercialairlines |
| Flight | Typeofaircraft |
| AirportFrom | Sourceairport |
| AirportTo | Destinationairport |
| DayOfWeek | Dayofthe week |
| Time | Departuretimemeasuredinminutesfrom  midnight(rangeisfrom10to1439) |
| Length | Durationoftheflightinminutes |
| Delay | Iftheflightisdelayed |

# DatasetSnapshot



**airports.xlsx**



# DatasetDescription



**airports.xlsx**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| id | Thisisanidentifierfortheairport.Itwillstaypersistentevenif  the airportcodechanges. |
| ident | Thisisthetextidentifierusedinthe*OurAirports*URL.Thiswillbethe International Civil Aviation Organization (ICAO) code ifavailable. Otherwise, it will be a local airport code (if there is noconflict)orwillbeaninternally-generatedcodestartingwiththe  ISO2countrycodefollowedbyadashandafour-digitnumber. |
| type | Thisshowsthe type oftheairport. Thevaluesallowed hereare  *closed\_airport*,*heliport*,*large\_airport*,*medium\_airport*,  *seaplane\_base*,and*small\_airport*. |
| name | Thisshowstheofficialnameoftheairport,including*Airport*and  *Airstrip* |
| latitude\_deg | Thisshowsthelatitudeof theairportin decimaldegrees(northis  positive). |
| longitude\_deg | Thisshowsthelongitudeoftheairportindecimaldegrees(east  ispositive). |

# DatasetDescription



**airports.xlsx**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| elevation\_ft | ThisshowstheelevationMSLoftheairportinfeet(notmeters). |
| continent | This shows the code for the continent where the airport is(primarily)located.Theallowedvaluesinclude*AF* (Africa),*AN*(Antarctica),*AS*(Asia),*EU*(Europe),*NA*(NorthAmerica),*OC*  (Oceania),or*SA*(SouthAmerica). |
| iso\_country | Thisshowsthetwo-characterISO3166:1-alpha2codeforthecountry wheretheairportis(primarily)located.Ahandfulof  unofficial,non-ISOcodesarealsoinuse,suchas*XK*forKosovo. |
| iso\_region | This is an alphanumeric code for the high-level administrativesubdivision of a country where the airport is primarily located(e.g., province and governorate) prefixed by the ISO2 countrycodeandahyphen.*OurAirports*usesISO3166:2codeswheneverpossible,preferringhigheradministrativelevels,butalso  includessome custom codes. |
| municipality | Thisshowstheprimarymunicipalitythattheairportserves(whenavailable). Notethatthis isnotnecessarilythe municipality  wheretheairportisphysicallylocated. |

# DatasetDescription



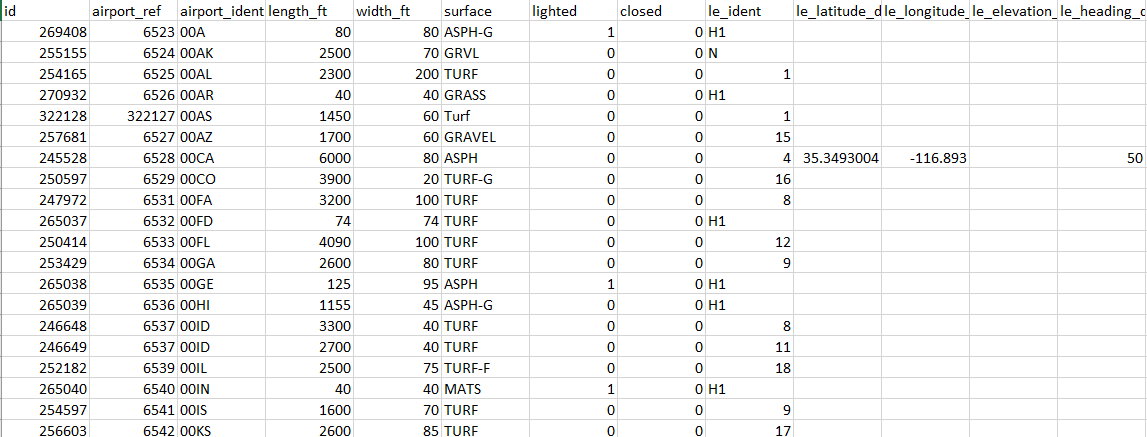
**airports.xlsx**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| scheduled\_service | Thisshows*yes*iftheairportcurrentlyhas scheduledairline  serviceand*no*ifotherwise. |
| gps\_code | ThisshowsthecodethatanaviationGPSdatabase(suchas,Jeppesen'sorGarmin's) wouldnormallyusefortheairport.  This will always be the ICAO code if one exists. Note that,unlikethe*ident*column,thisisnotguaranteedtobeglobally  unique. |
| iata\_code | Thisshowsthethree-letterIATAcodefortheairport(ifithas  one). |
| local\_code | Thisshowsthelocalcountrycodefortheairportifit’sdifferent  fromthe*gps\_code*and*iata\_code*fields(usedmainlyforUS  airports). |
| home\_link | ThisshowstheURLoftheairport'sofficialhomepageon the  webifoneexists. |
| wikipedia\_link | ThisshowstheURLoftheairport'spage onWikipediaifone  exists. |
| Keywords | This field contains other keywords or phrases to assist with thesearch. These are separated by a comma. It may also includeformernamesfortheairport,alternatecodes,namesinother  languages,and nearbytouristdestinations. |

# DatasetSnapshot



**runways.xlsx**



# DatasetDescription



**runways.xlsx**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| id | This showstheinternal*OurAirports*integeridentifierfortherunway.  Thiswillstaypersistentevenif therunway numberingchanges. |
| airport\_ref | Thisshowstheinternalintegerforeignkeymatchingthe*id*columnfortheassociatedairportin **airports.csv**.Here,*airport\_ident*isa  betteralternative. |
| airport\_ident | Thisshowstheexternally-visiblestringforeignkeymatching  the*ident*columnfortheassociatedairportin**airports.csv**. |
| length\_ft | Thisshowsthelengthofthefullrunwaysurface(includingdisplaced  thresholdsandoverrunareas)infeet. |
| width\_ft | Thisshows thewidth of therunwaysurfaceinfeet. |
| surface | This shows the code for the runway surface type. This is not acontrolledvocabularyyetbutprobablywillbesoon.Somecommonvaluesinclude*ASP*(asphalt),*TURF*(turf),*CON*(concrete),*GRS*(grass),  *GRE*(gravel),*WATER*(water),and*UNK*(unknown). |

# DatasetDescription



**runways.xlsx**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| lighted | Thisshows*1*ifthesurfaceislitatnightand*0*ifnot.Notethatthisisinconsistent with**airports.csv**whichuses*yes*and*no*  instead.) |
| closed | Thisshows*1*iftherunwaysurfaceiscurrentlyclosed and*0*if  not. |
| le\_ident | Thisshowstheidentifierforthelow-numberedendofthe  runway. |
| le\_latitude\_deg | Thisshowsthelatitudeofthecenterofthelow-numbered  endof therunwayindecimaldegrees(northispositive)if  available. |
| le\_longitude\_deg | Thisshowsthelongitudeofthecentreofthelow-numberedendoftherunwayindecimaldegrees(eastispositive)if  available. |
| le\_elevation\_ft | ThisshowstheelevationaboveMSLofthelow-numberedend  oftherunwayinfeet. |
| le\_heading\_degT | Thisshowstheheadingofthelow-numberedendofthe  runwayin degreestrue(notmagnetic). |

# DatasetDescription



**runways.xlsx**

|  |  |
| --- | --- |
| **Variables** | **Description** |
| le\_displaced\_threshold\_ft | Thisshowsthelengthofthedisplacedthreshold(ifany)for  thelow-numberedendoftherunwayinfeet. |
| he\_ident | Thisshowstheidentifierforthehigh-numberedendofthe  runway. |
| he\_latitude\_deg | Thisshowsthelatitudeofthecentreofthehigh-numberedendoftherunwayindecimaldegrees(northispositive)if  available. |
| he\_longitude\_deg | Thisshowsthelongitudeofthecentreofthehigh-  numberedendoftherunwayindecimaldegrees(eastis  positive)ifavailable. |
| he\_elevation\_ft | Thisshowsthe elevationaboveMSLofthe high-numbered  endoftherunway infeet. |
| he\_heading\_degT | Thisshowstheheadingofthehigh-numberedendofthe  runwayin degreestrue(notmagnetic). |
| he\_displaced\_threshold\_ft | Thisshowsthelengthofthedisplacedthreshold(ifany)for  thehigh-numberedendoftherunwayinfeet. |

# ProjectTask:Week1



**ApplieddatasciencewithPython**

1. Importandaggregatedata:
   1. Collect information related to flights, airports (e.g., type of airport and elevation), and runways(e.g.,*length\_ft*,*width\_ft*,*surface*,andnumberofrunways).Gatherallfieldsyoubelievemightcauseavoidabledelaysinonedataset.

Hint: In this case, you would have to determine the keys to join the tables. A data description will beuseful.

* 1. Whenitcomestoon-timearrivals,differentairlinesperformdifferentlybasedontheamountofexperience they have. The major airlines in this field include US Airways Express (founded in1967) Continental Airlines (founded in 1934), and Express Jet (founded in 19860. Pull suchinformation specific to various airlines from the Wikipedia page link given below.<https://en.wikipedia.org/wiki/List_of_airlines_of_the_United_States>.

Hint: Here,youshoulduse webscrapingto learnhowlonganairlinehasbeenoperating.

# ProjectTask:Week1



**ApplieddatasciencewithPython**

* 1. Youshouldthengetalltheinformationgatheredsofarinone place.
  2. The totalpassengertrafficmay alsocontributetoflightdelays.Theterm*hub*referstobusycommercialairports.*Largehubs*areairportsthataccountforatleast1percentof thetotal passengerenplanementsintheUnitedStates.Airportsthataccountfor

0.25percentto1percentoftotalpassengerenplanementsareconsideredmediumhubs. Pull passenger traffic data from the Wikipedia page given below using webscrapingandcollateitinatable.

<https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States>

1. You should then examine the missing values in each field, perform missing valuetreatment,andjustifyyouractions.

# ProjectTask:Week1



**ApplieddatasciencewithPython**

1. Performdata visualizationandshareyourinsightsonthefollowingpoints:
   1. Accordingtothe dataprovided,approximately70%of SouthwestAirlinesflightsare

delayed.Visualizeittocompareitwiththedataofotherairlines.

* 1. Flightsweredelayedonvariousweekdays.Whichdayoftheweekis thesafestfortravel?
  2. Whichairlinesshouldberecommendedforshort-,medium-,andlong-distancetravel?
  3. Doyounoticeanypatternsinthedeparturetimesoflong-durationflights?

1. Howmanyflightsweredelayedatlargehubscomparedtomediumhubs?Useappropriatevisualizationtorepresentyourfindings.

# ProjectTask:Week1



**ApplieddatasciencewithPython**

1. Usehypothesistestingstrategiestodiscover:
   1. Iftheairport'saltitudehasanythingtodowithflightdelaysforincominganddepartingflights
   2. Ifthenumberof runwaysat anairportaffectsflight delays
   3. Ifthedurationofaflight(length)affectsflightdelays

Hint:Testthisfromtheperspectiveofboththesourceanddestinationairports

1. Findthecorrelationmatrixbetweentheflightdelaypredictors,createaheatmapto

visualizethis,andshareyourfindings

# ProjectTask:Week1



**Machinelearning**

1. UseOneHotEncoderandOrdinalEncodertodealwithcategoricalvariables
2. Performthefollowingmodel building steps:
   1. Applylogisticregression(usestochasticgradientdescentoptimizer)anddecisiontree

models

* 1. Usethestratified five-foldmethodtobuildandvalidatethemodels

**Note**:Makesureyouusestandardizationeffectively,ensuringnodataleakageandleveragepipelinestohaveacleanercode

* 1. UseRandomizedSearchCVforhyperparametertuning,andusek-foldforcross-validation
  2. Keepafewdatapoints(10%)forpredictionpurposestoevaluatehowyouwould

makethefinalprediction,anddonotusethisdatafortestingorvalidation

# ProjectTask:Week1



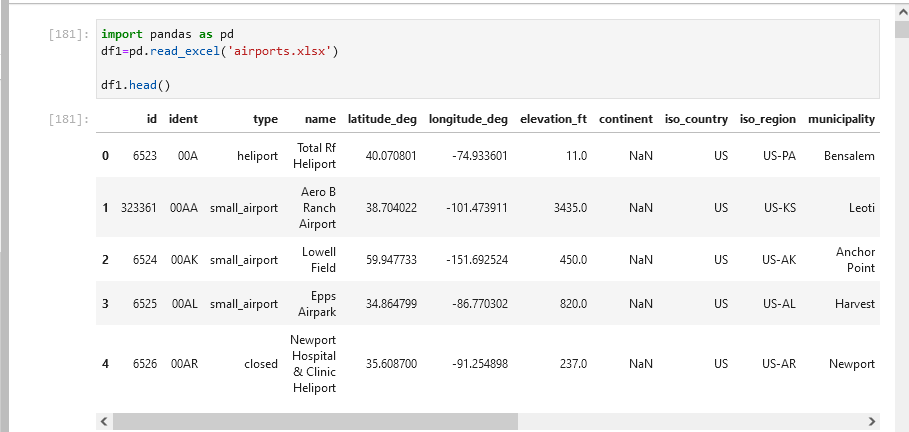
**Machinelearning**

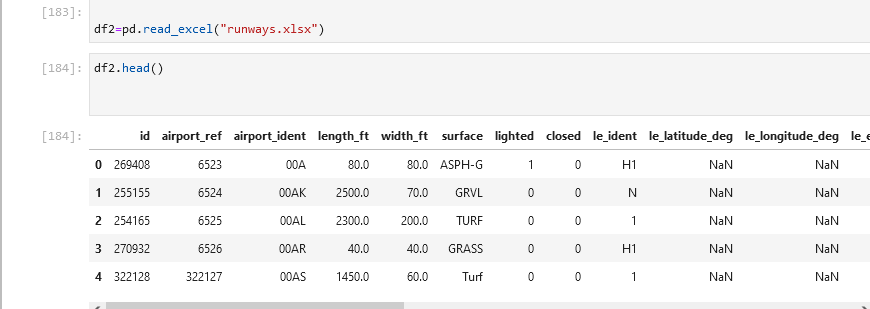
**Note:**Thefinalprediction willbebasedonthevoting(majorityclassby5modelscreated

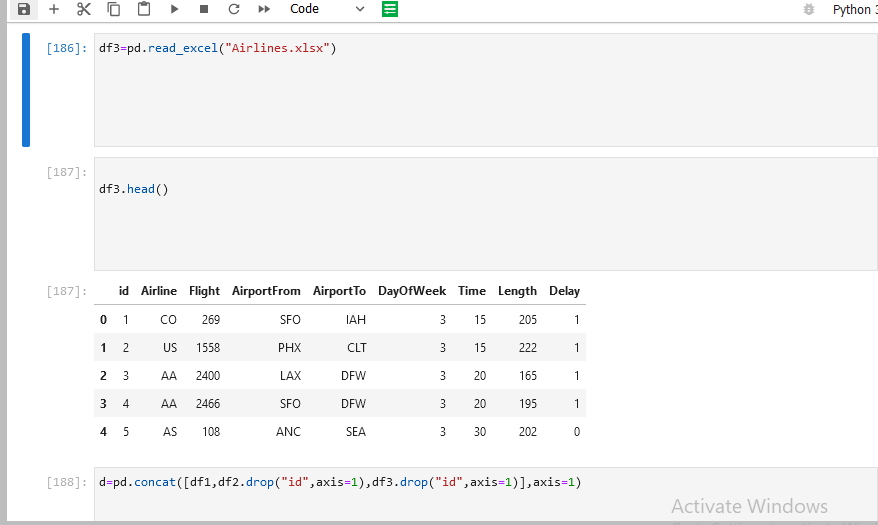
usingthestratified5-foldmethod)

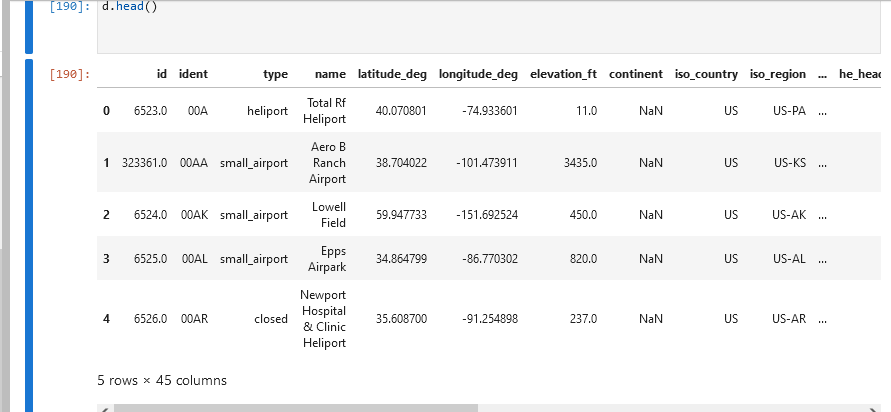
g.Comparetheresultsoflogisticregressionanddecisiontreeclassifier

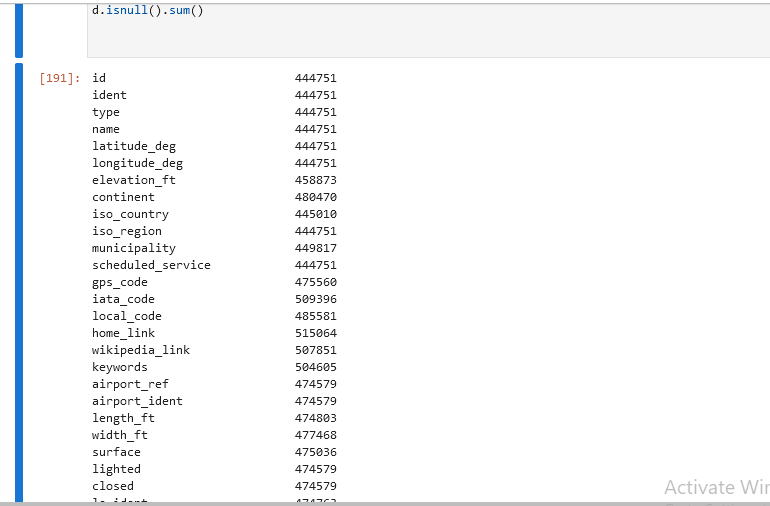
1. Use the stratified five-fold method to build and validate the models using the XGBclassifier,compareallmethods,andshareyourfindings

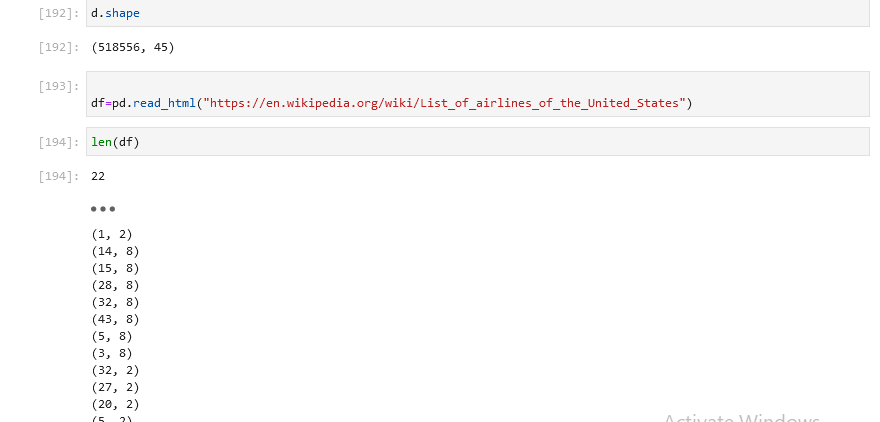


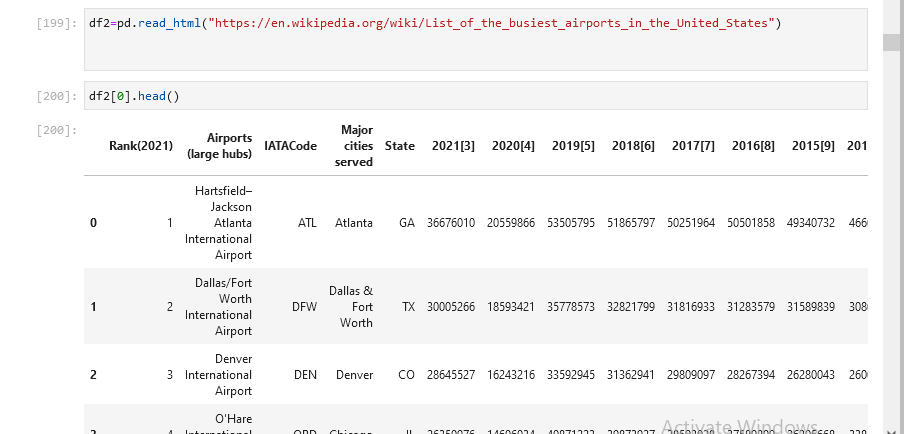




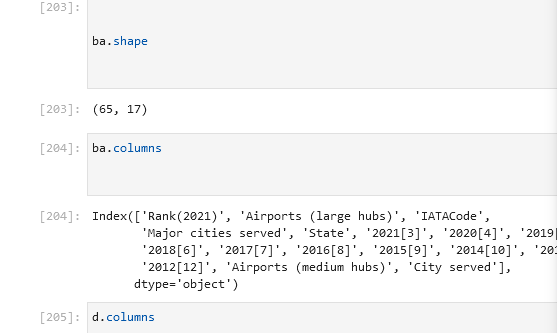


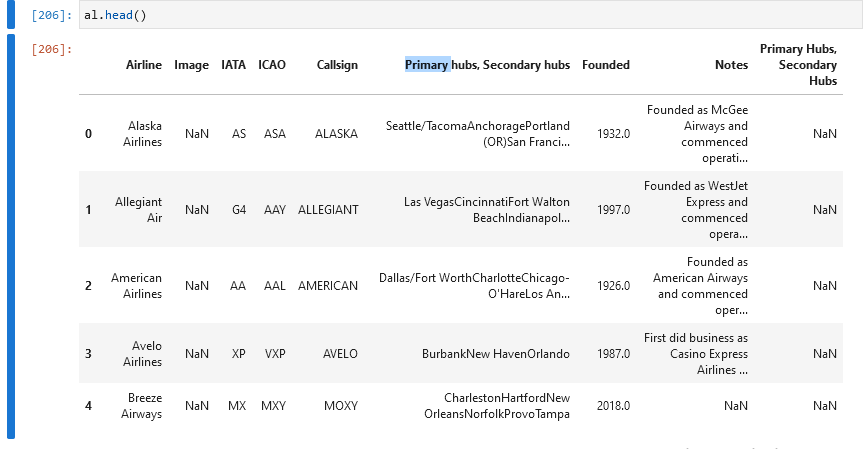


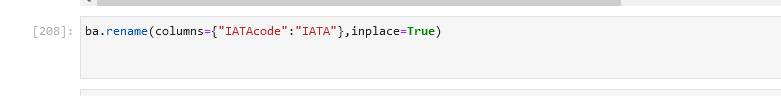






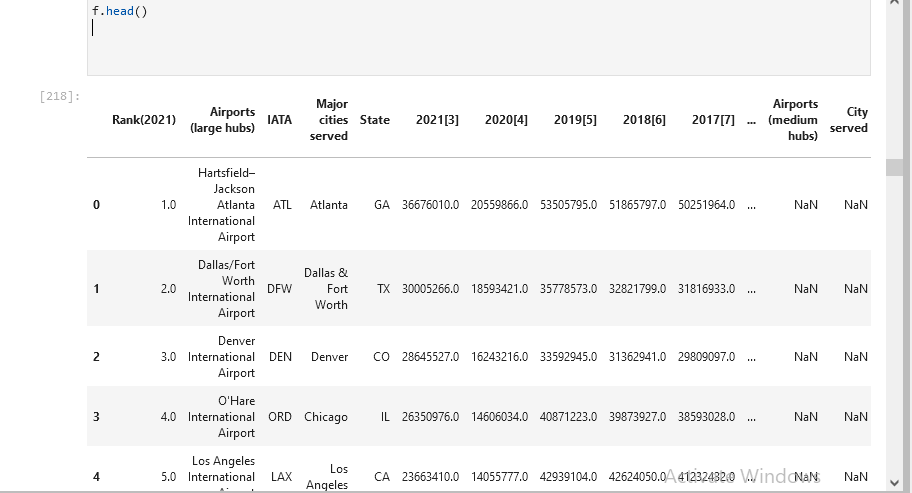


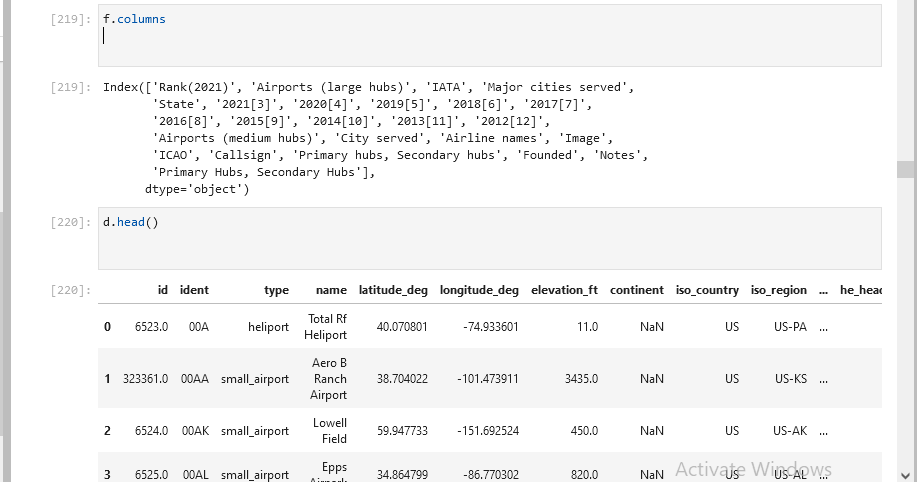




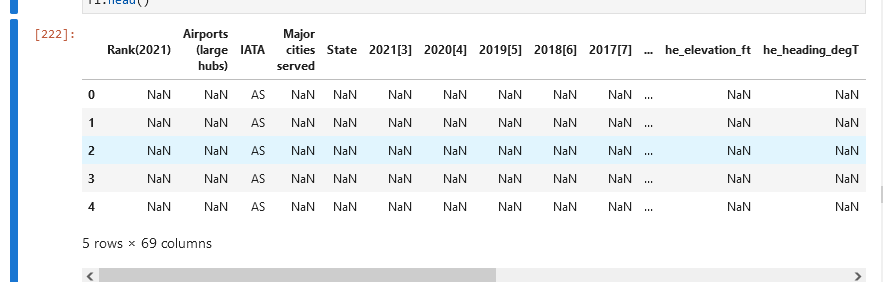




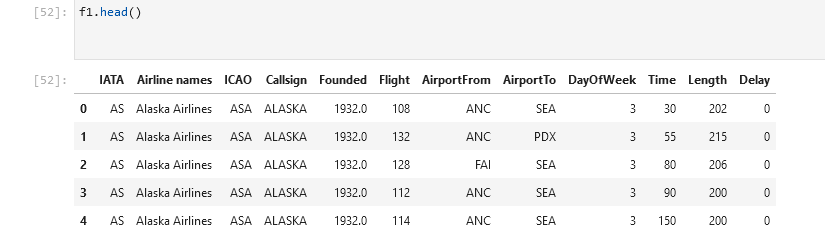


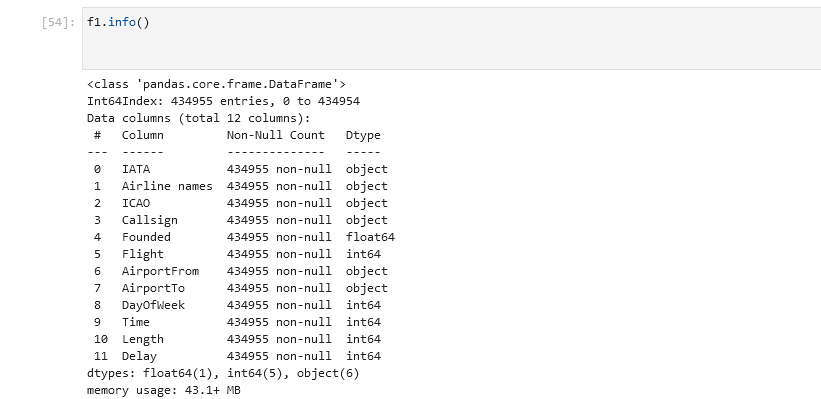


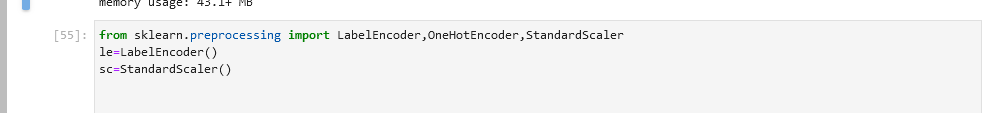


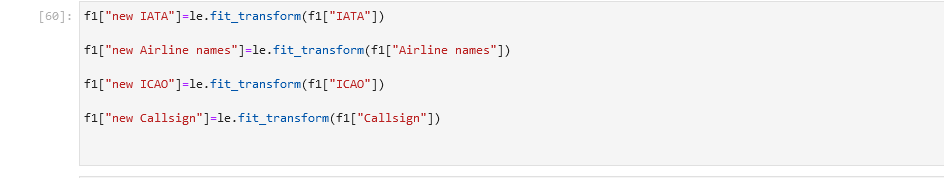


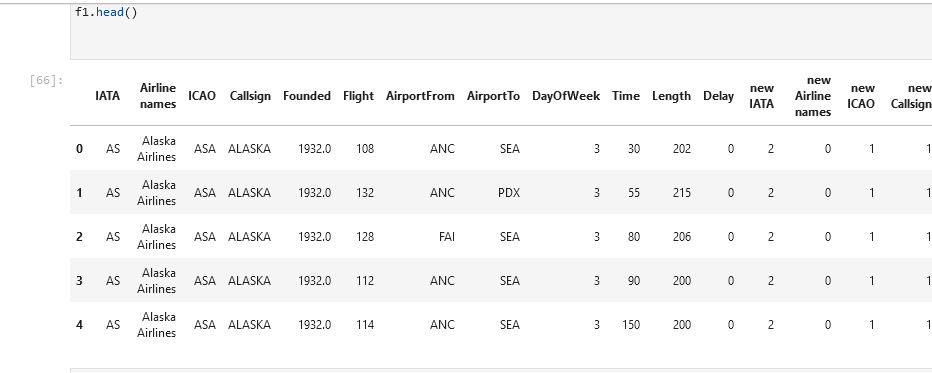




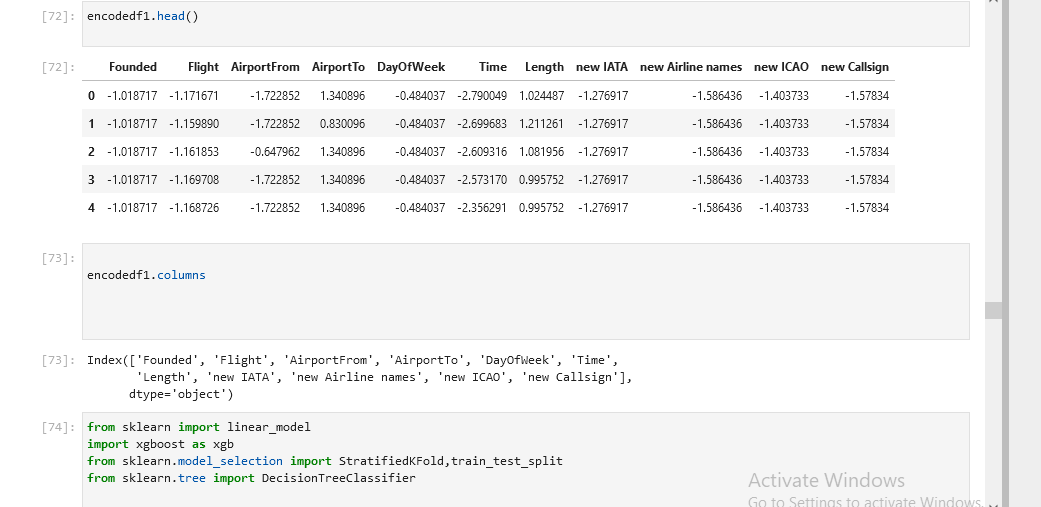


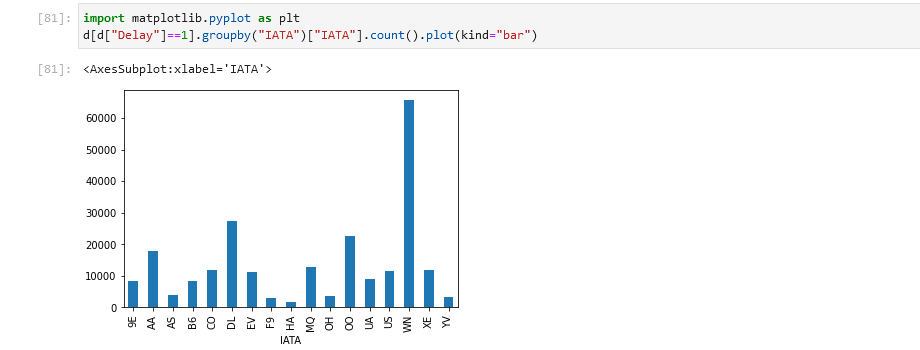


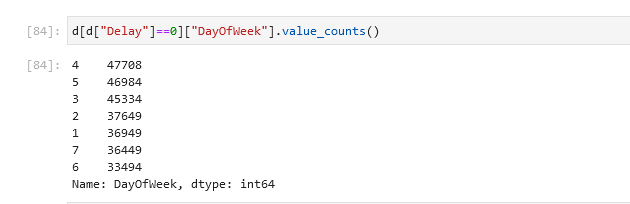




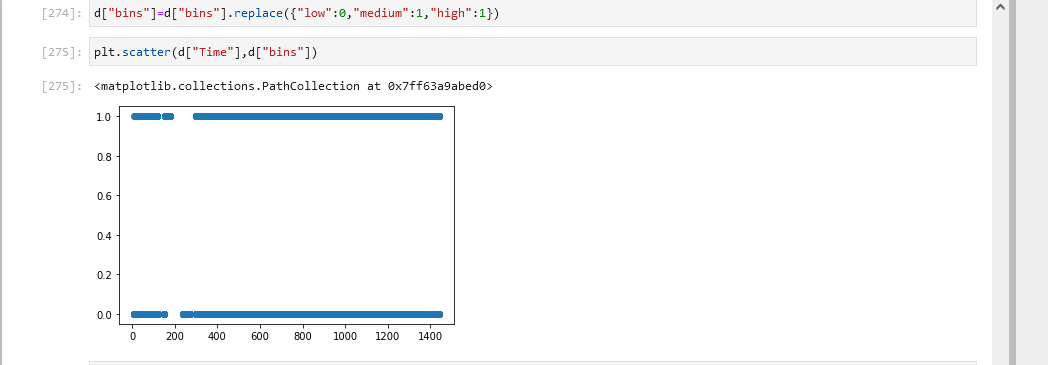


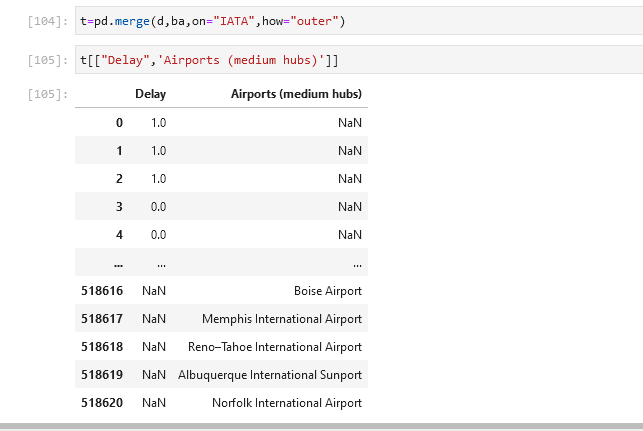


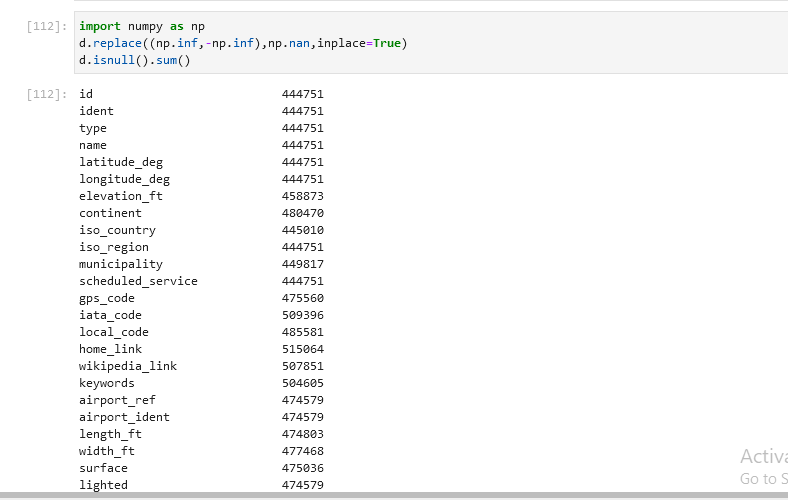




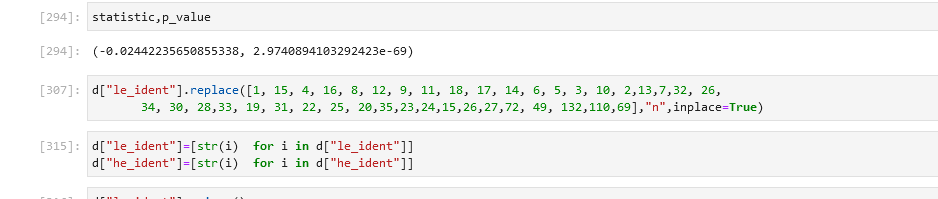






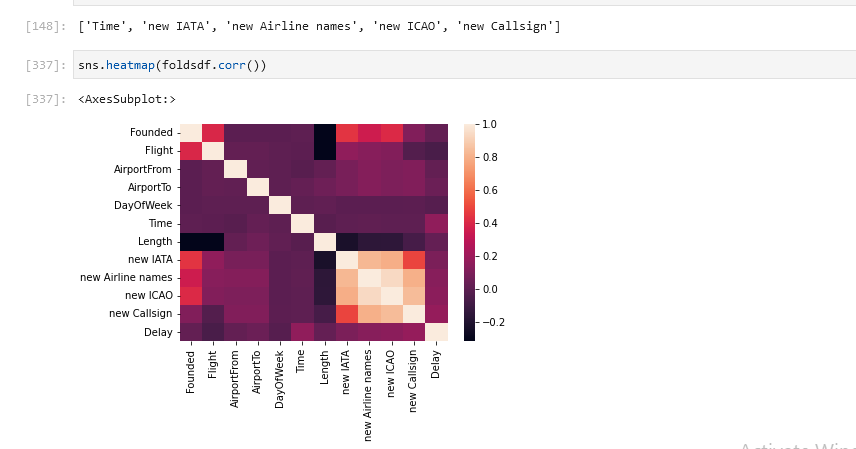














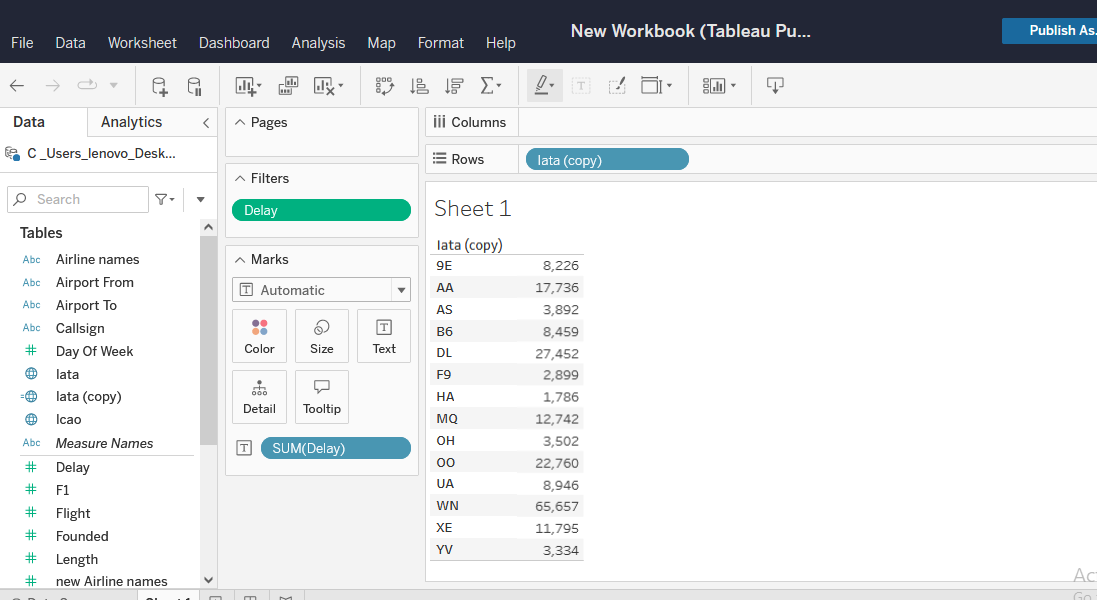
# ProjectTask:Week2

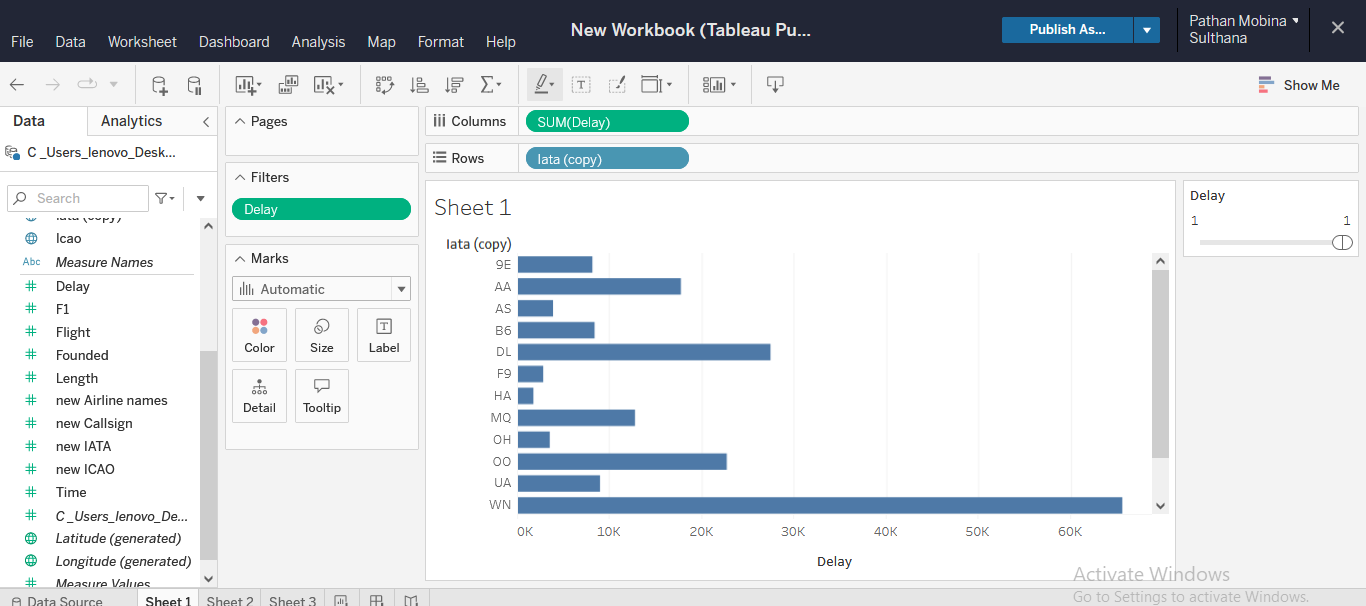


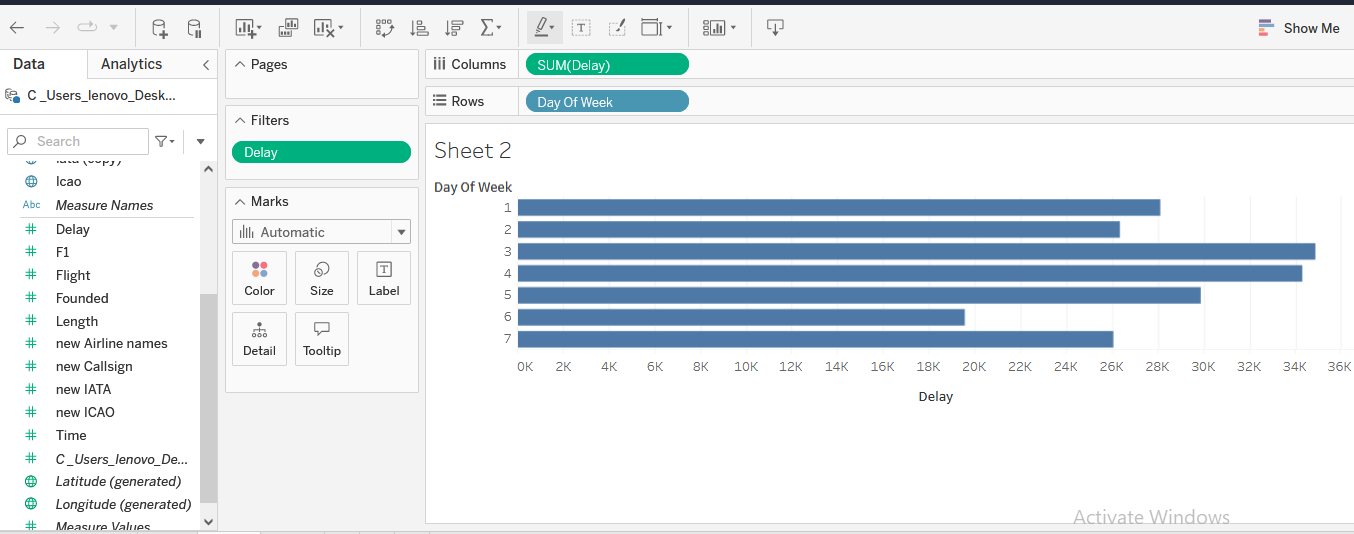
**Tableau**

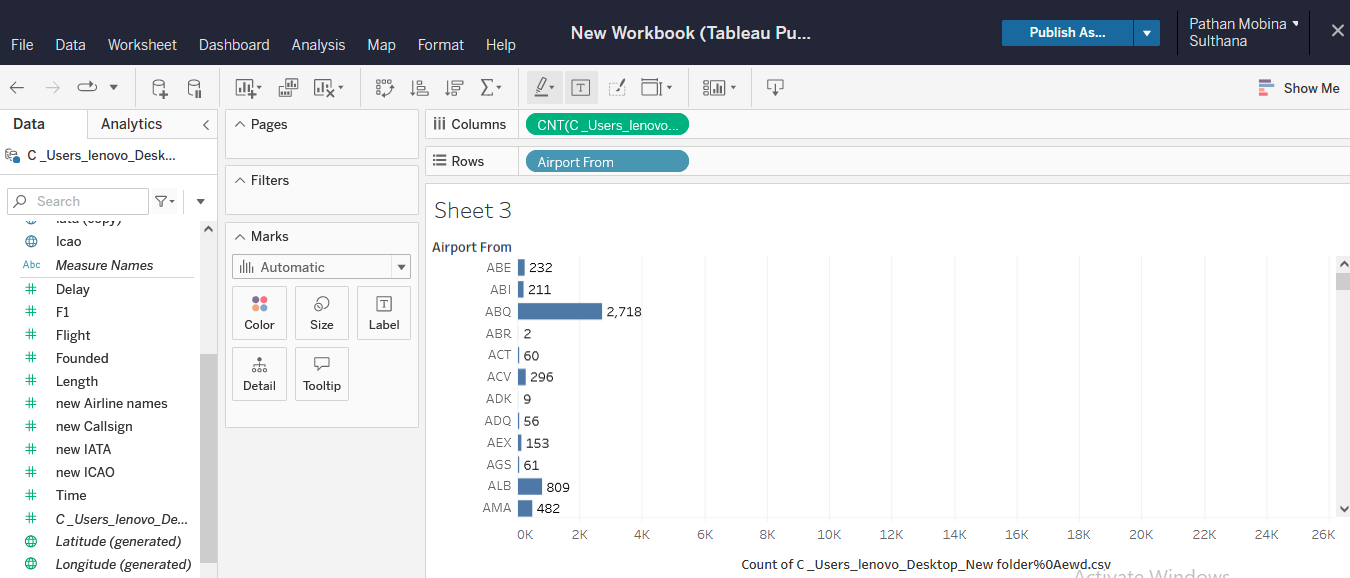
1. CreateadashboardinTableaubyselectingappropriatecharttypesandmetricsforthebusiness

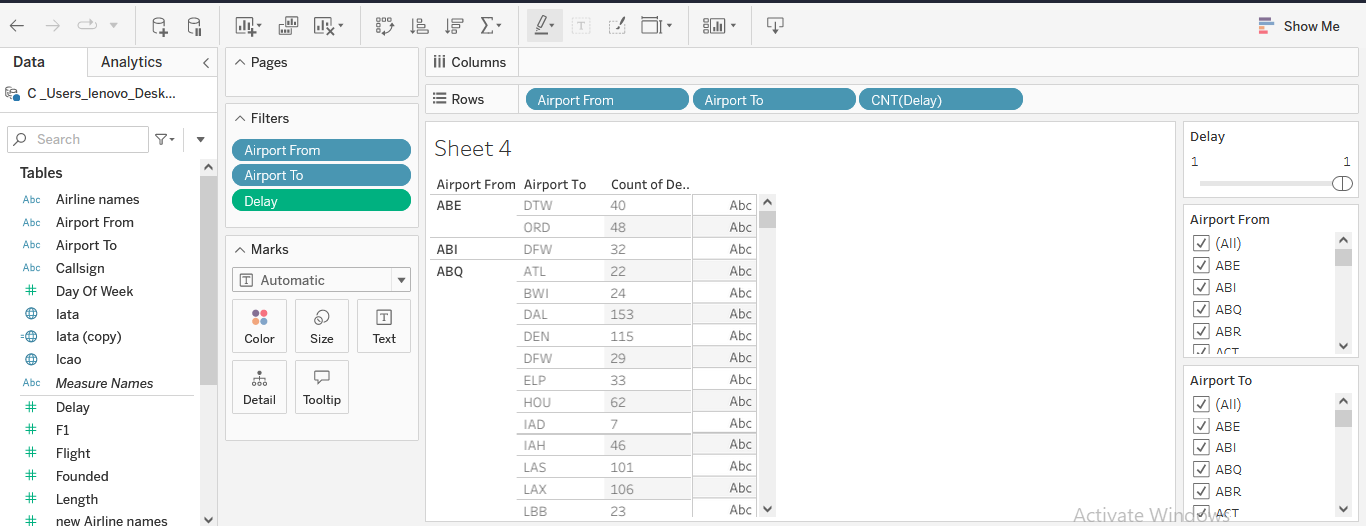
**Note:**Putmoreemphasisondatastorytelling



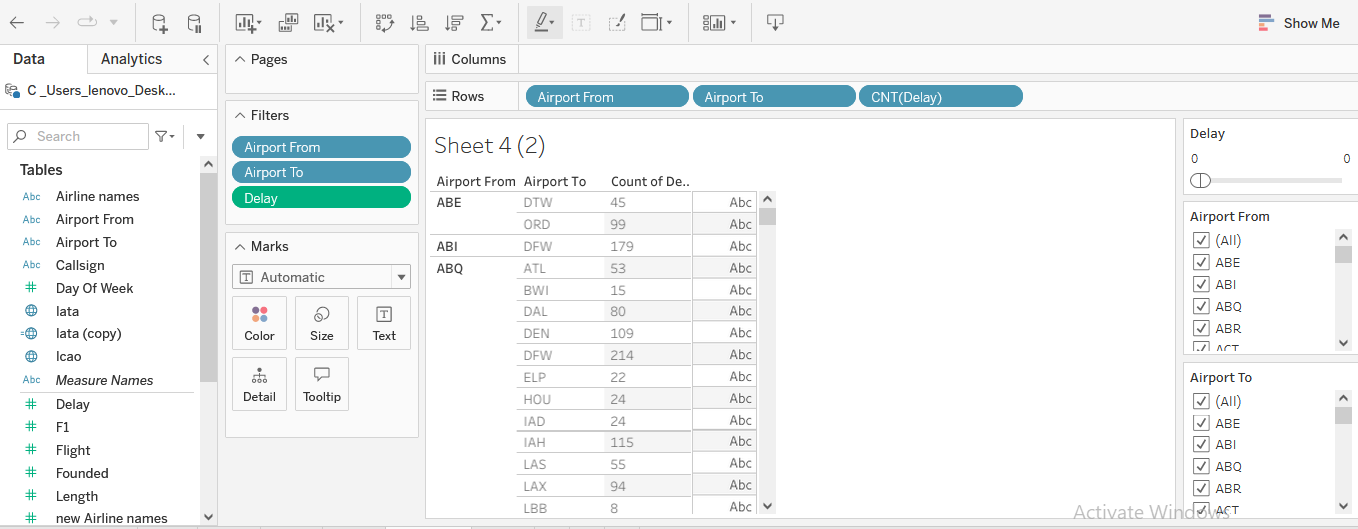


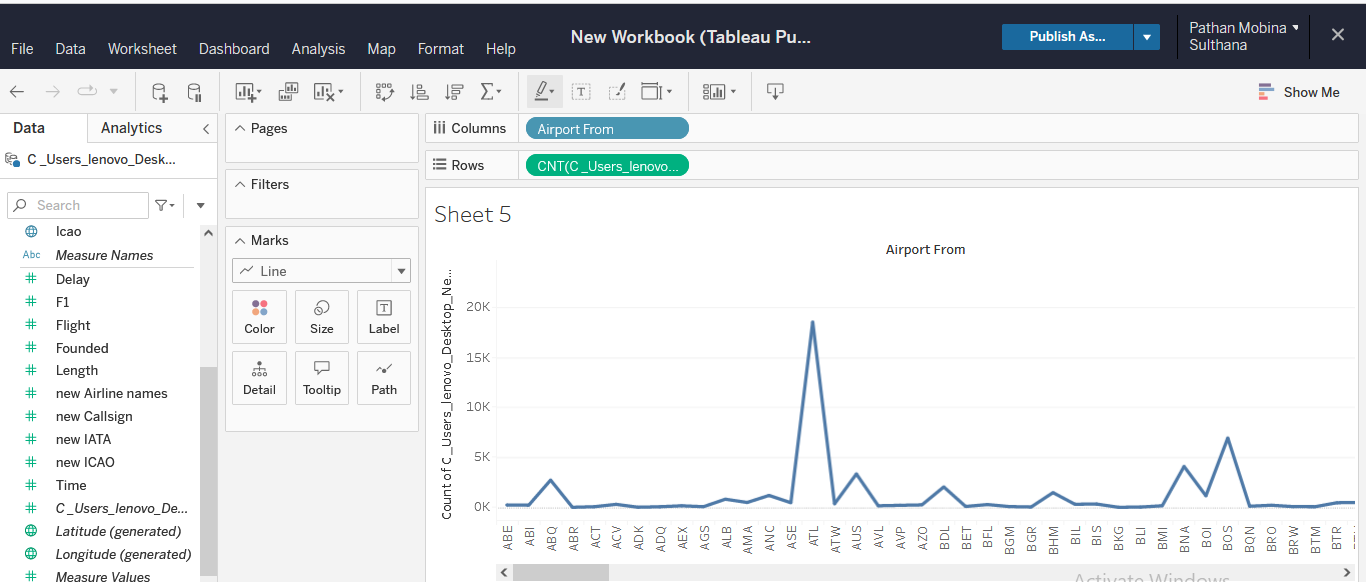


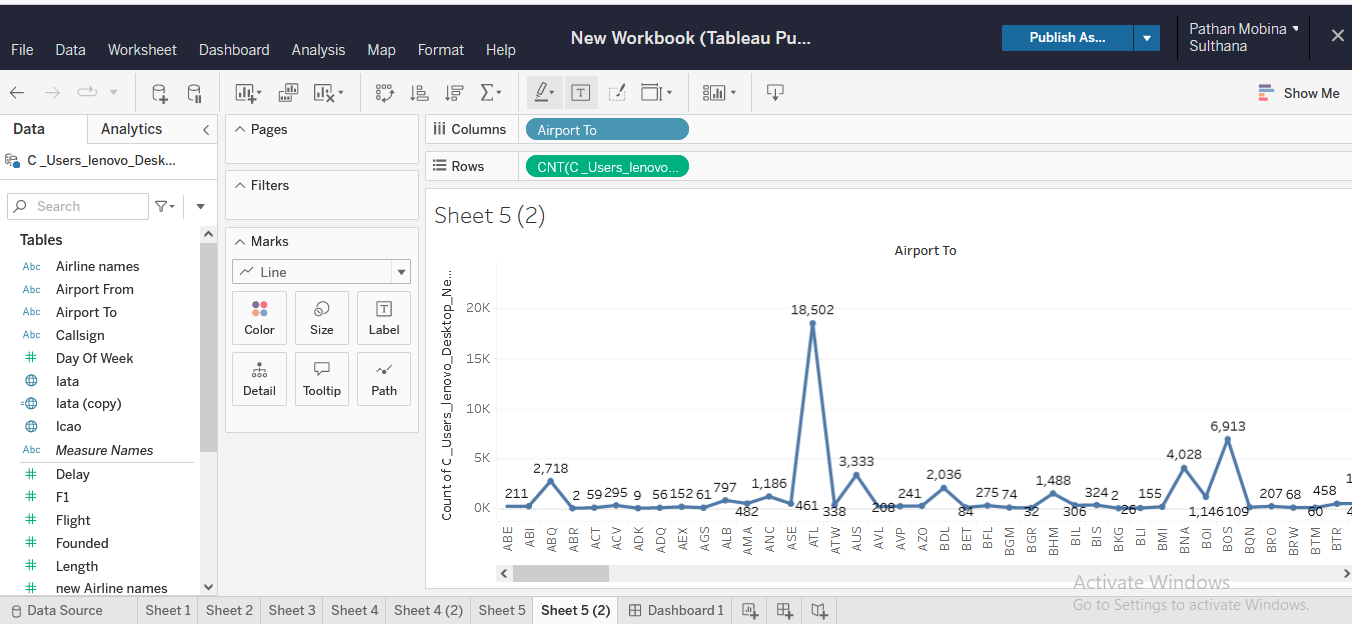


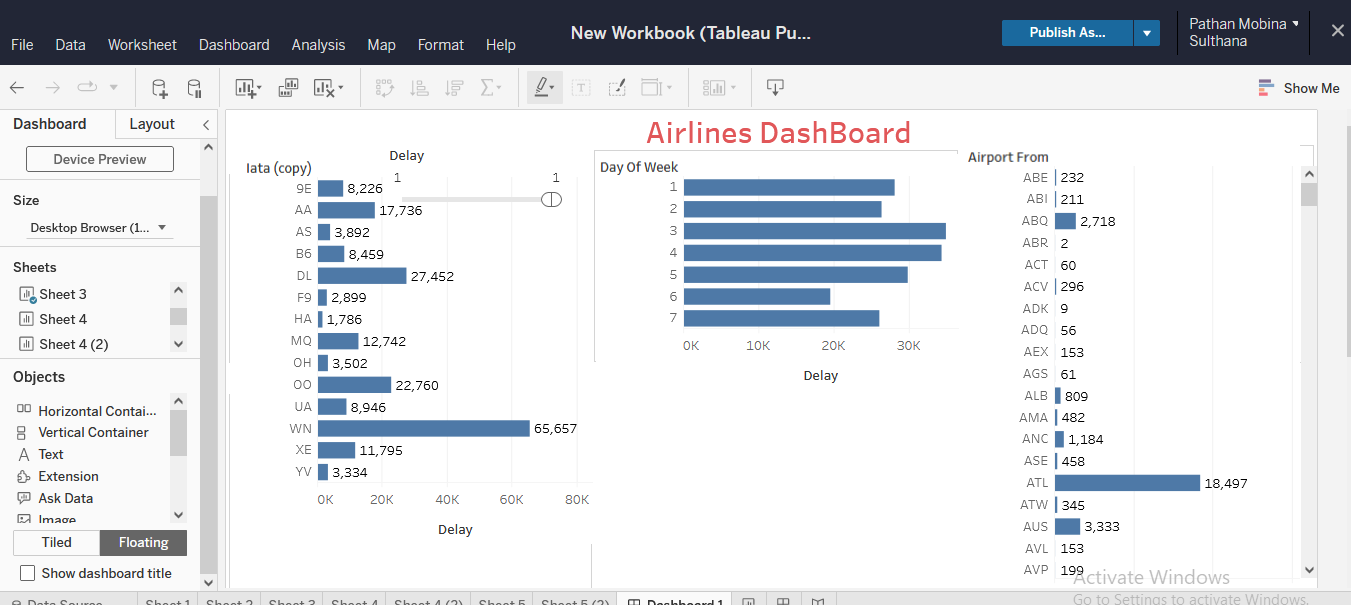


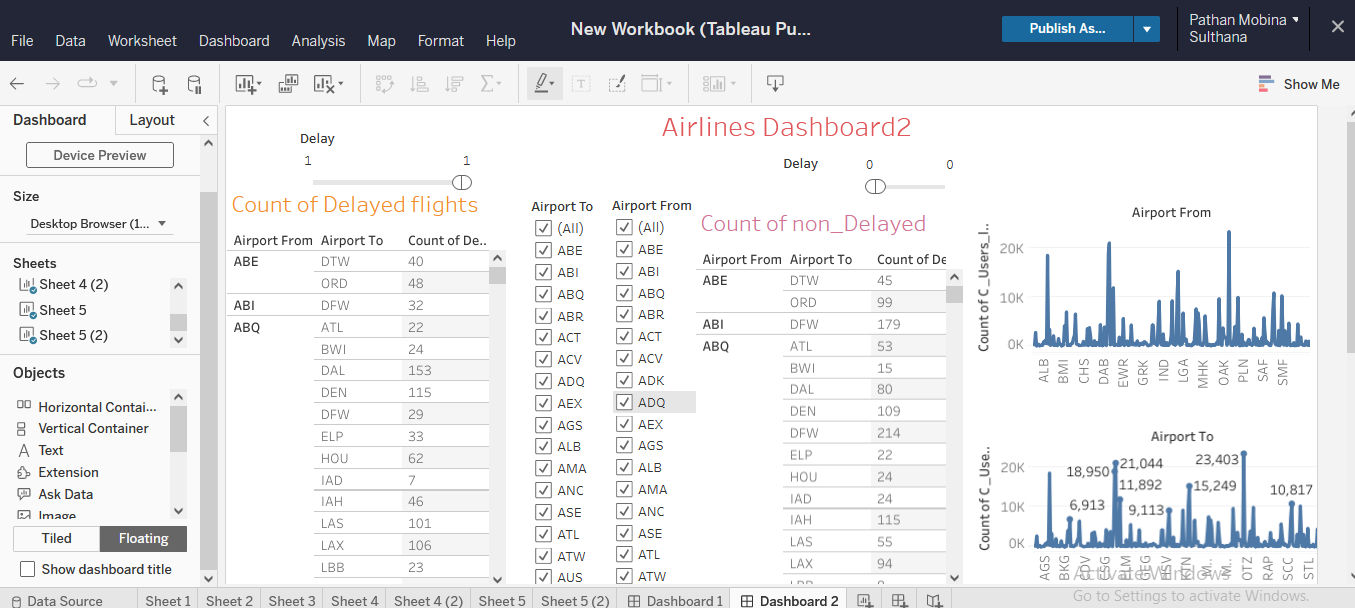
\











# ProjectTask:Week2

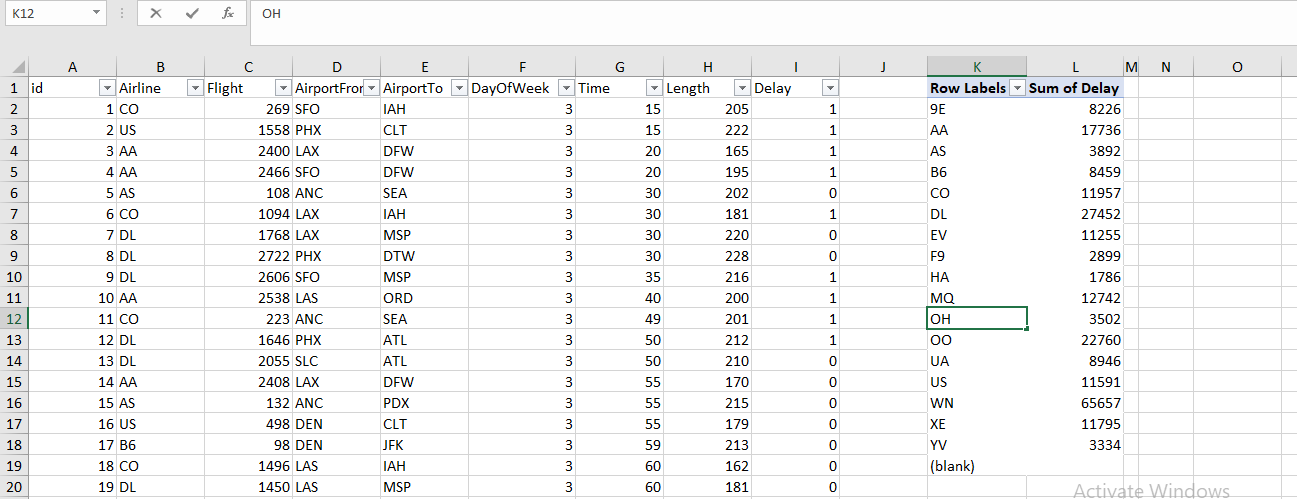


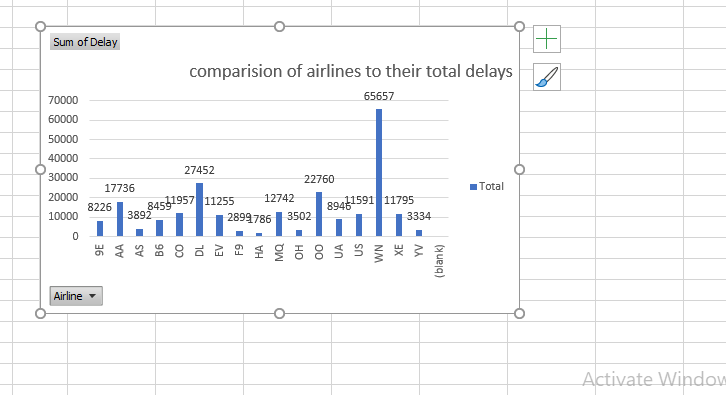
**Excel**

1. CreateanExceldashboardshowcasingthefollowing(useformcontrolsto makeadynamic

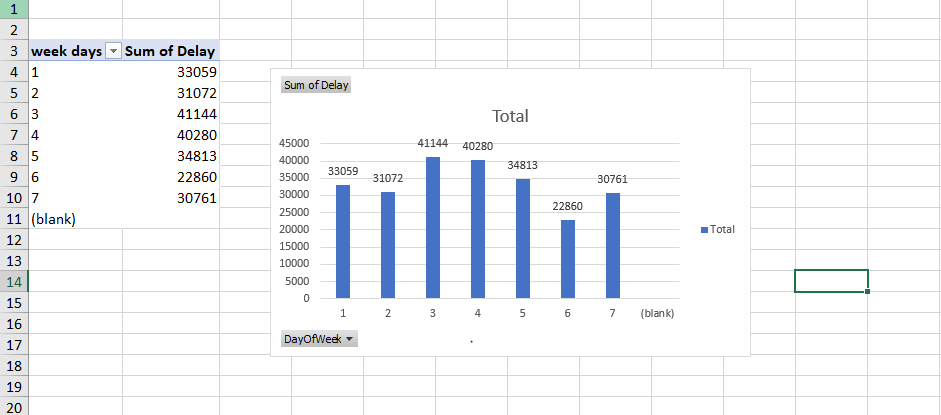
chart):

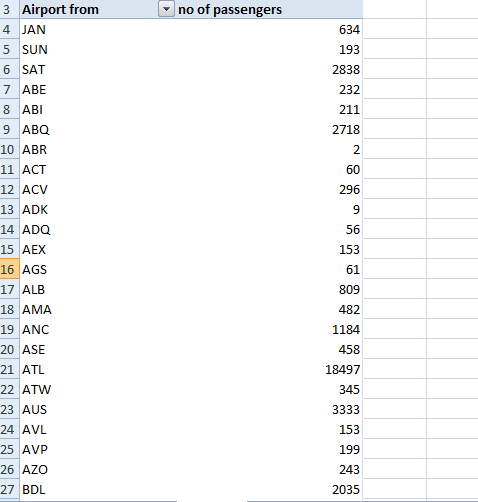
* 1. Comparedifferentairlinesbasedontheiron-timeperformance

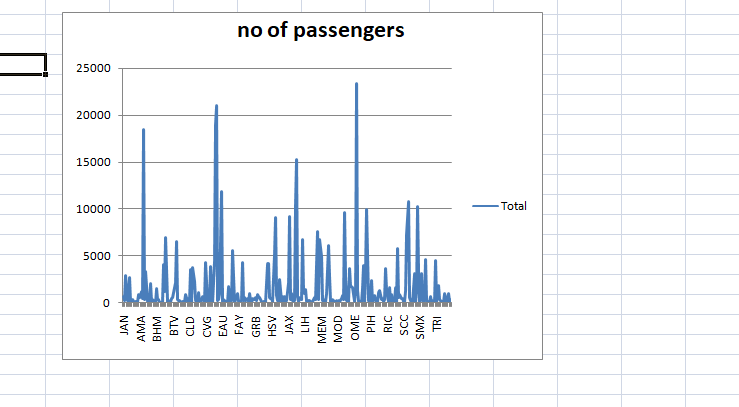




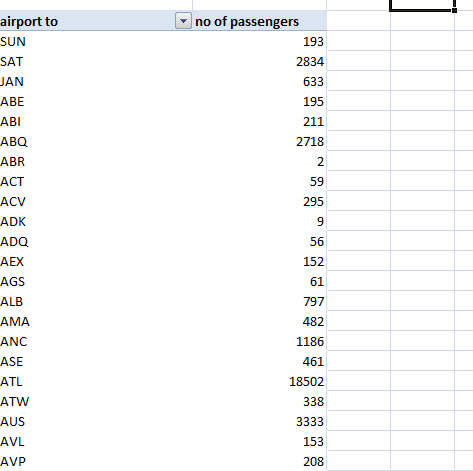
* 1. Comparethepercentageofdelayedflightsfordifferentdaysoftheweek

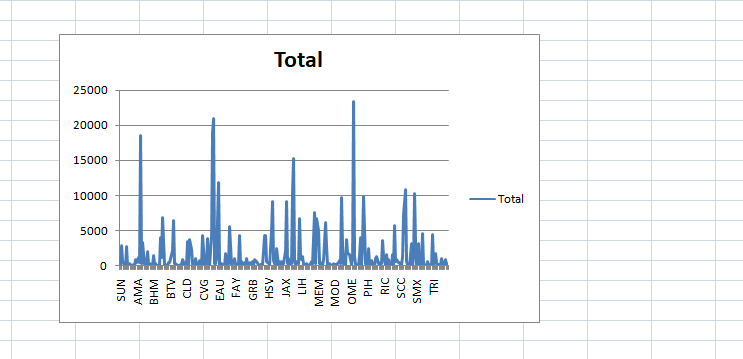


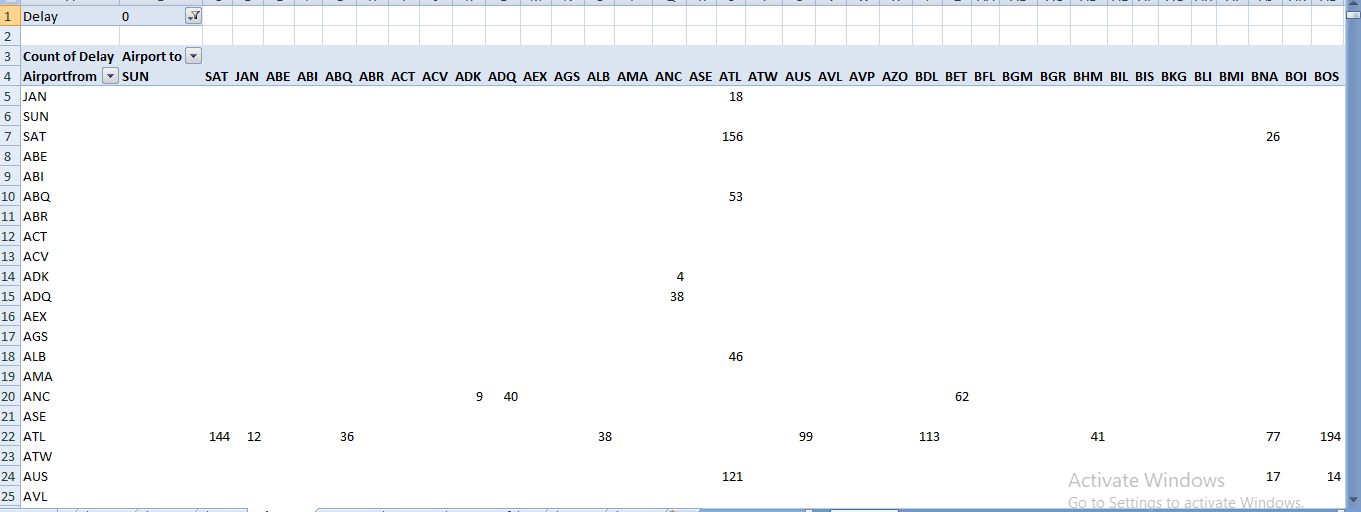


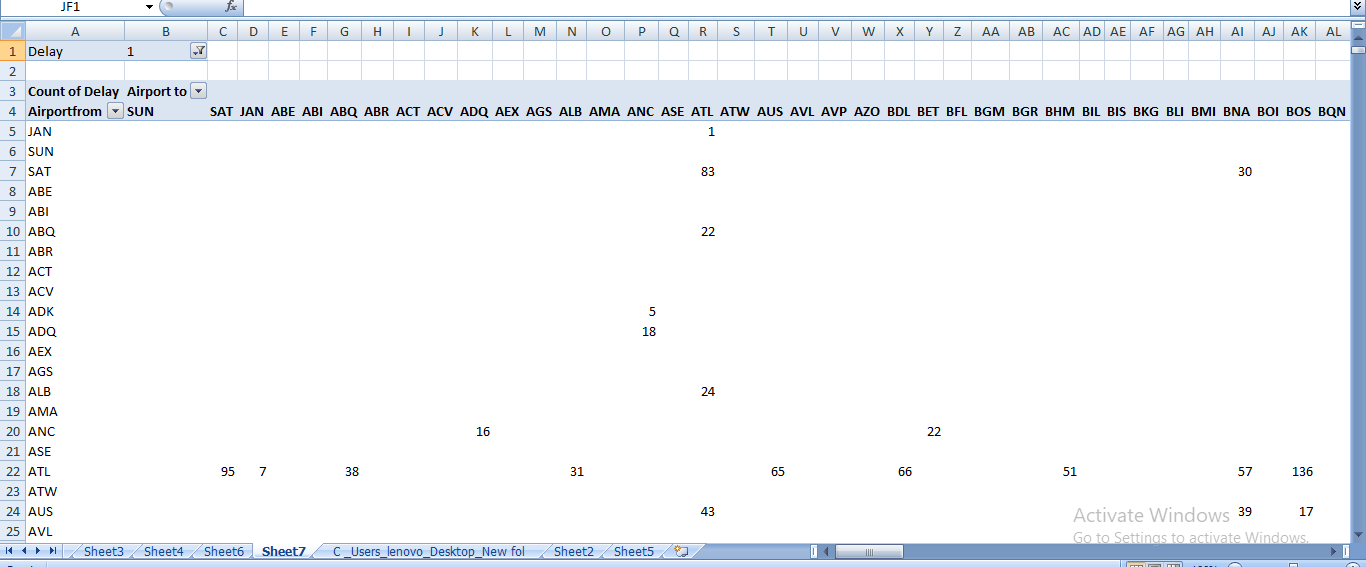


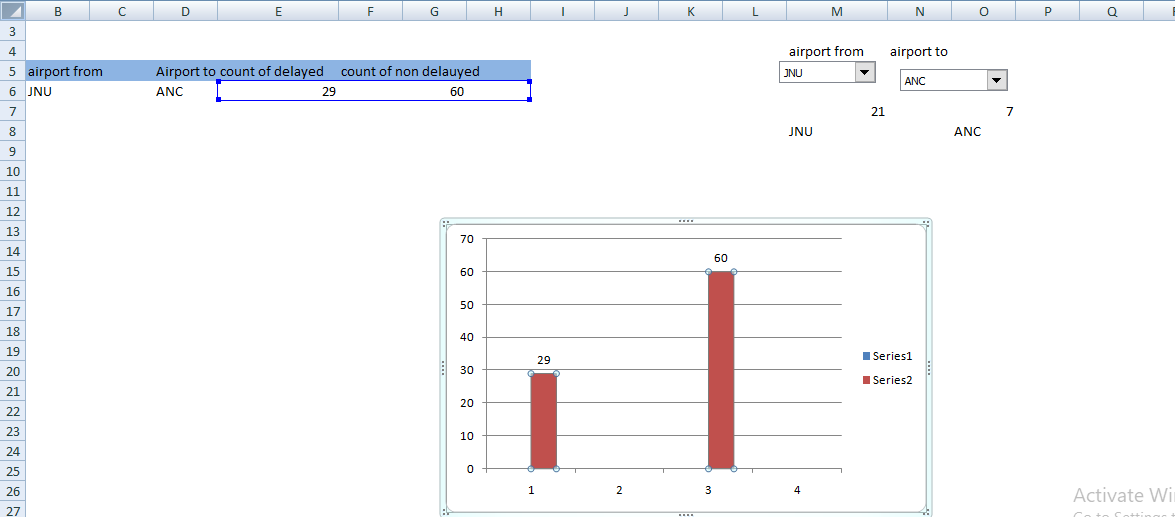
* 1. Createatrendchartforthenumberofpassengersatlargeandmediumhubs
  2. Visualizethecountofdelayedandon-timeflightsfordifferentpairsofsourceanddestinationairports
     + Createadynamicchartthatallowsuserstoselectasourceanddestinationairport.

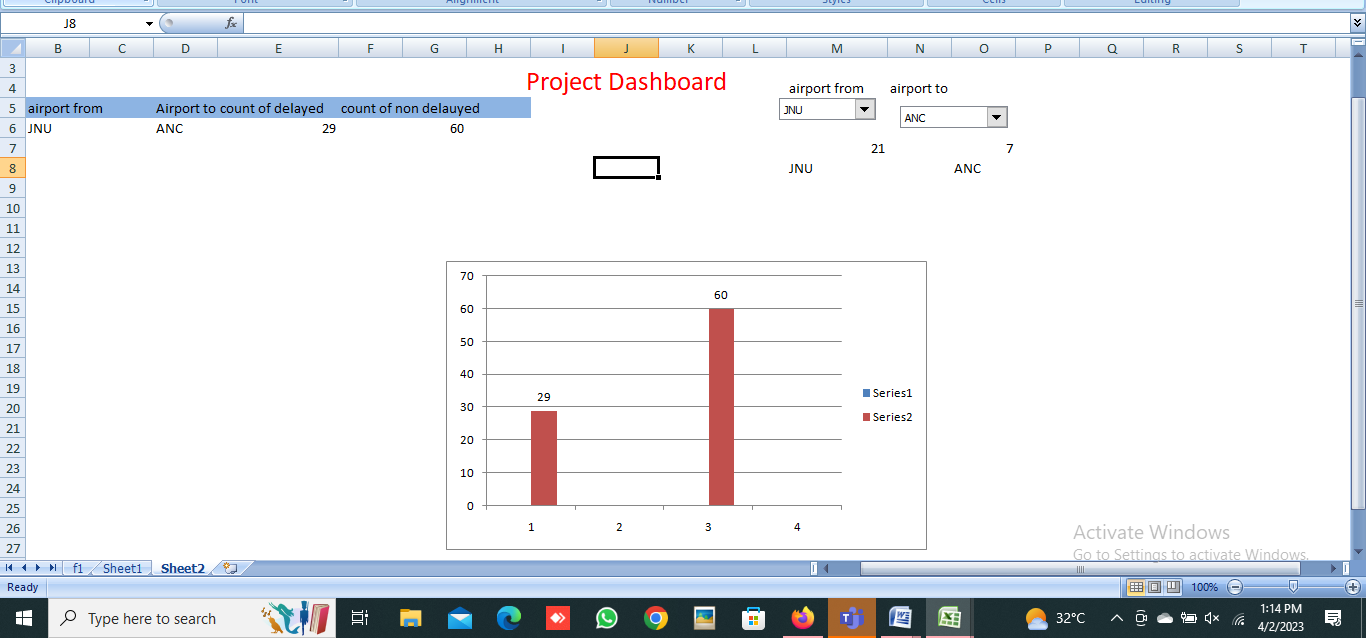












# ProjectTask:Week2



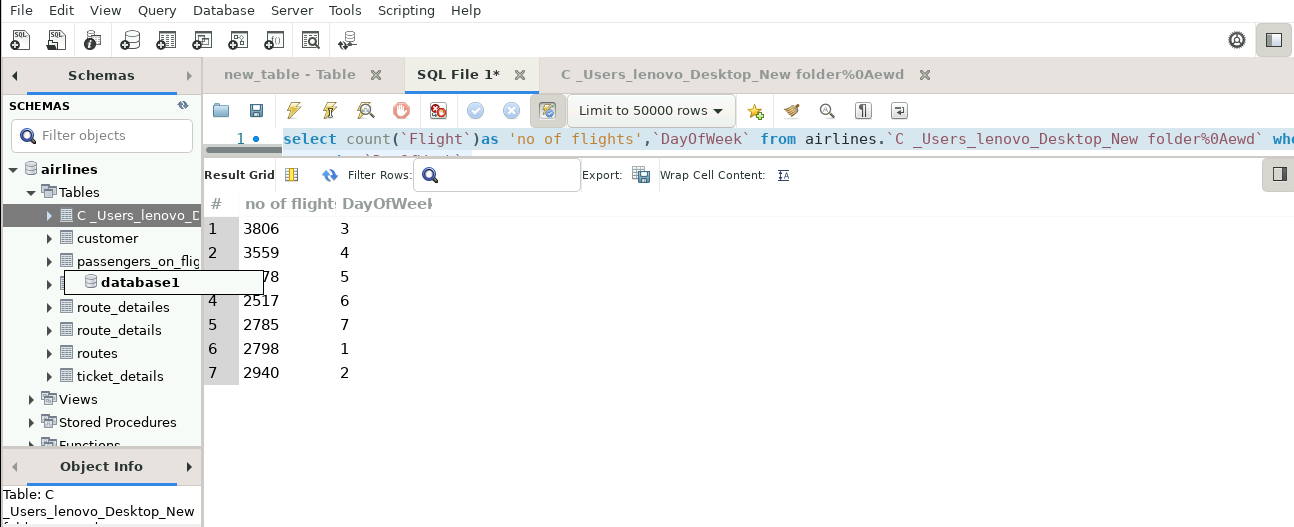
**SQL**

1. Determine the number of flights that are delayed on various days of the week.

select count(`Flight`)as 'no of flights',`DayOfWeek` from airlines.`C \_

Users\_lenovo\_Desktop\_New folder%0Aewd` where `Delay`=1

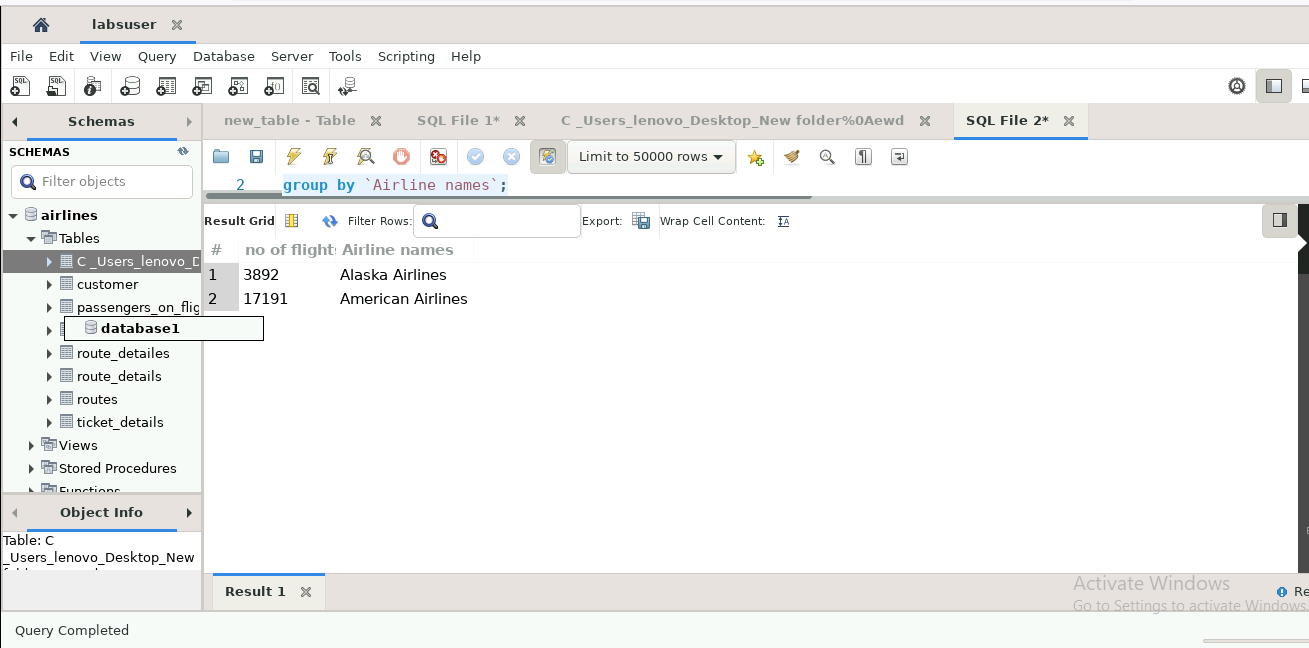
group by `DayOfWeek`;



1. Determine the number of delayed flights for various airlines.

select count(`Flight`)as 'no of flights',`Airline names` from airlines.`C \_Users\_lenovo\_Desktop\_New folder%0Aewd` where `Delay`=1

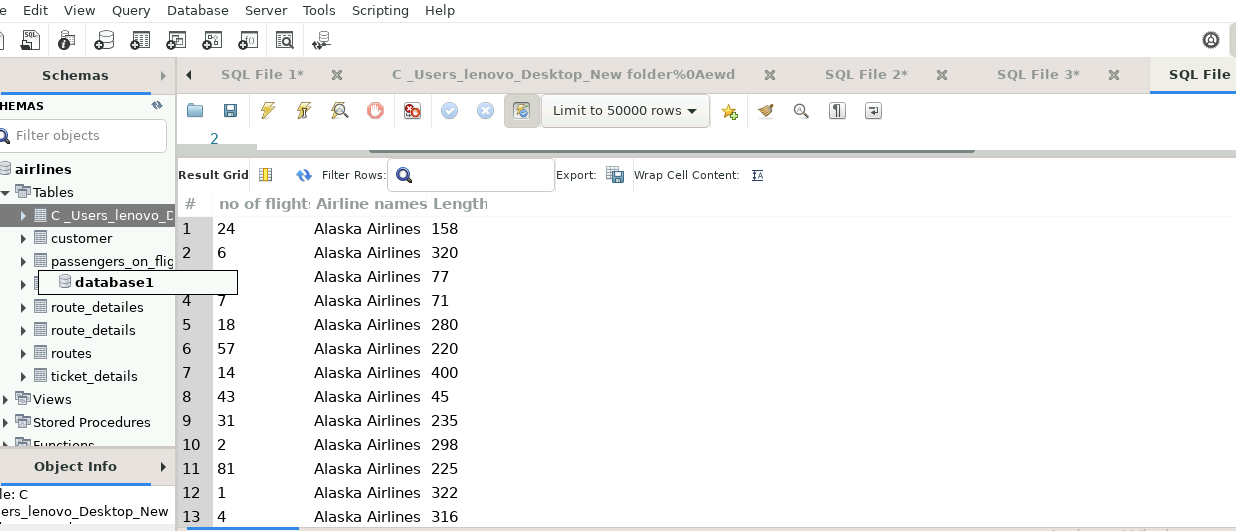
group by `Airline names`;



1. Determine how many delayed flights land at airports with atleast 10 runways.

select count(`Flight`)as 'no of flights',`Airline names`,`Length` from airlines.`C \_Users\_lenovo\_Desktop\_New folder%0Aewd` where `Delay`=1 and `Length`>=10

group by `Airline names`,`Length`;

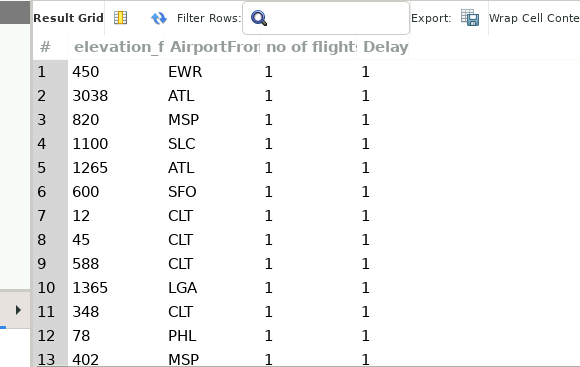


1. Compare the number of delayed flights at airports higher than average elevation and

Those that are lower than average elevation for both source and destination airports

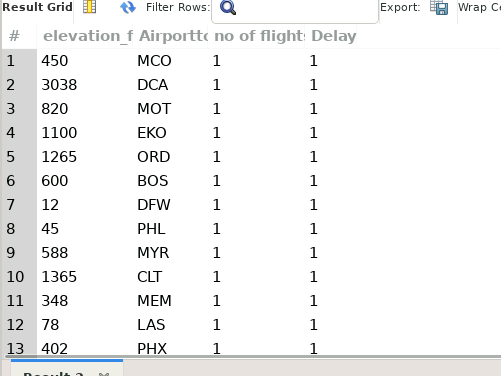
select elevation\_ft ,AirportFrom ,count(Flight)as'no of flights',Delay from airports1

join Airlines1 as l on airports1.id=l.id where Delay=1 group by elevation\_ft,AirportFrom;



select elevation\_ft ,Airportto ,count(Flight)as'no of flights',Delay from airports1

join Airlines1 as l on airports1.id=l.id where Delay=1 group by elevation\_ft,Airportto;



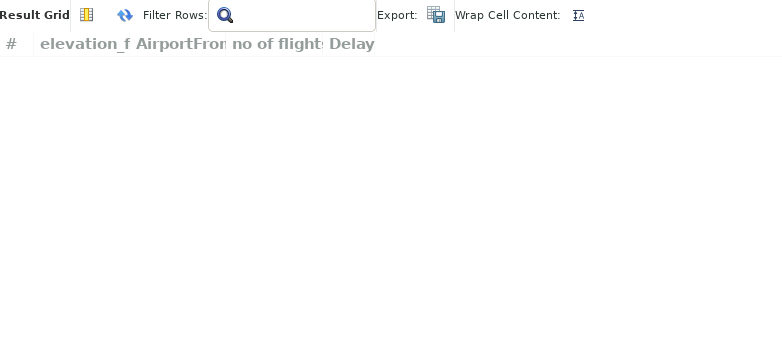
select elevation\_ft ,Airportto ,count(Flight)as'no of flights',Delay from airports1

join Airlines1 as l on airports1.id=l.id where Delay=1 group by elevation\_ft,Airportto having elevation\_ft>avg(elevation\_ft);



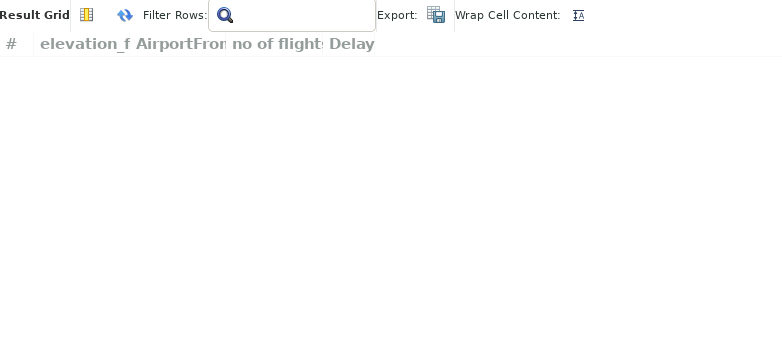
select elevation\_ft ,AirportFrom ,count(Flight)as'no of flights',Delay from airports1

join Airlines1 as l on airports1.id=l.id where Delay=1 group by elevation\_ft,AirportFrom having elevation\_ft>avg(elevation\_ft);



select elevation\_ft ,AirportFrom ,count(Flight)as'no of flights',Delay from airports1

join Airlines1 as l on airports1.id=l.id where Delay=1 group by elevation\_ft,AirportFrom having elevation\_ft<avg(elevation\_ft);

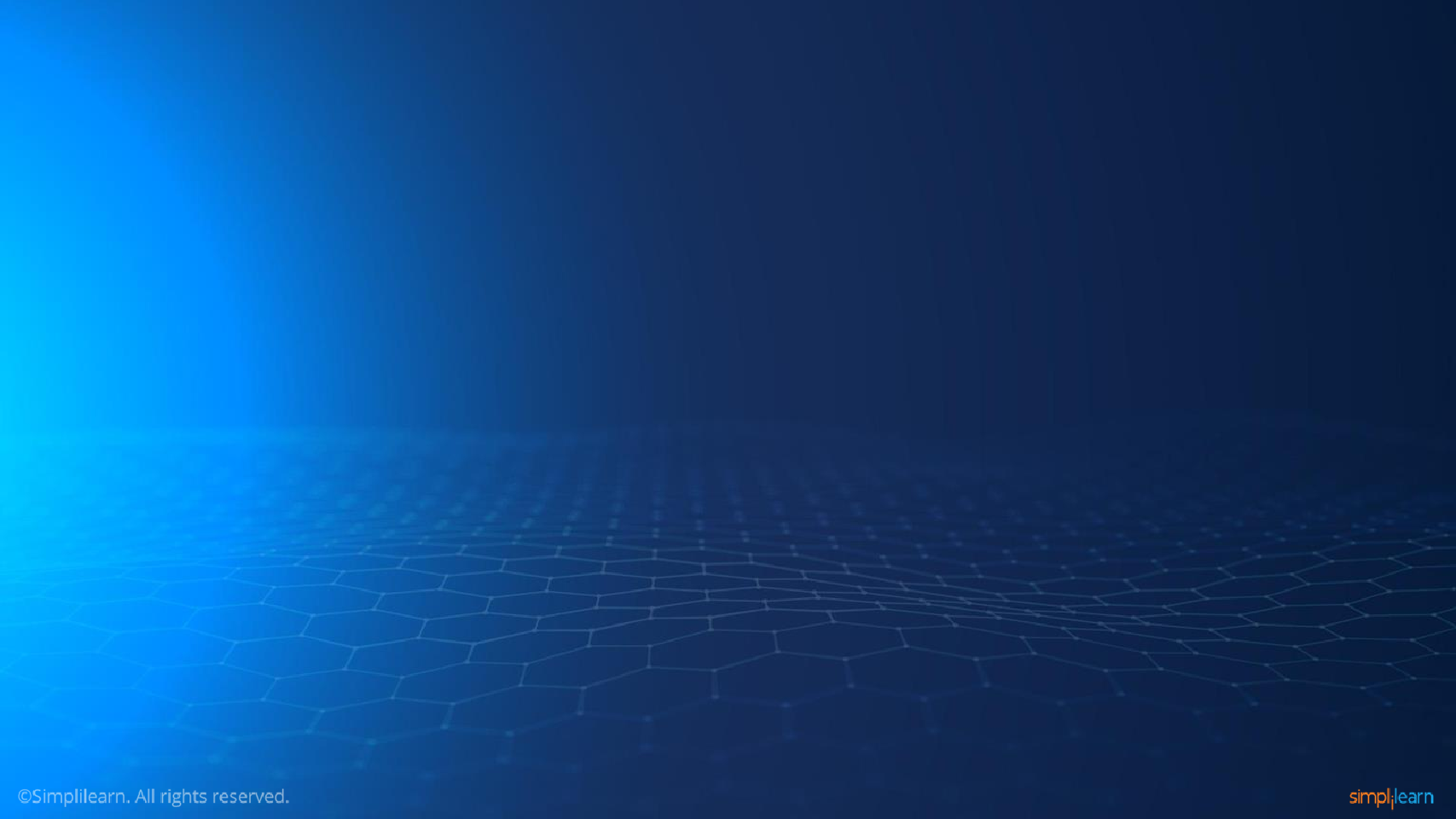


select elevation\_ft ,Airportto ,count(Flight)as'no of flights',Delay from airports1

join Airlines1 as l on airports1.id=l.id where Delay=1 group by elevation\_ft,Airportto having elevation\_ft<avg(elevation\_ft);







**ThankYou**