**0**

**ITM 818 Data Management and Visualization**

Homework 2: Data Management (60 Points)

1. Please use the flightdata.csv dataset (which was used in Lab4) to answer the following questions in R (Notice: do not use the “sqldf” package). (20 Points)

(a). Get the subset of flights with distances greater than the average value of distances and sort the data by distance in descending order. Please return the first 5 records of the output.

(b). Get the subset of flights that departure delay is not missing and sort the subset data by departure delay in descending order. Please return the first 10 records of the sorted subset data (Hint: be careful about difference between NA and NULL).

(c). Find all the flights between “4/5/2016” and “4/10/2016” (including these two days). Please return the first 6 records of the output.

(d). Calculate cancellation rate of each carrier (i.e. a cancellation rate of 0.05 should mean that 5% of flights are cancelled), and order by the rate (descending). Please return a table with two columns, namely the rate and name of carrier.

(e). Add one categorical variable using departure delay based on the following rules: “NoDelay” if delay time is negative, zero, or missing; “LowDelay” if delay time is greater than 0 and smaller than or equals to median of positive departure delay values; and “HighDelay” if delay time is greater than median of positive departure delay values. Return the first 10 records of the new data.

**CODE**

df = read.csv("C:/Users/neilr/OneDrive/Documents/R/DataSet/flightdata.csv")

head(df)

library(dplyr)

# Q1. ------------------------

a = df %>% filter(distance > mean(distance)) %>% arrange(desc(distance))

head(a,5)

b = df %>% subset(!is.na(df$depdelay)) %>% arrange(desc(depdelay))

head(b,10)

c = df %>% filter(as.Date(date) >= "2016-04-01" & as.Date(date) <= "2016-04-10" )

head(c,6)

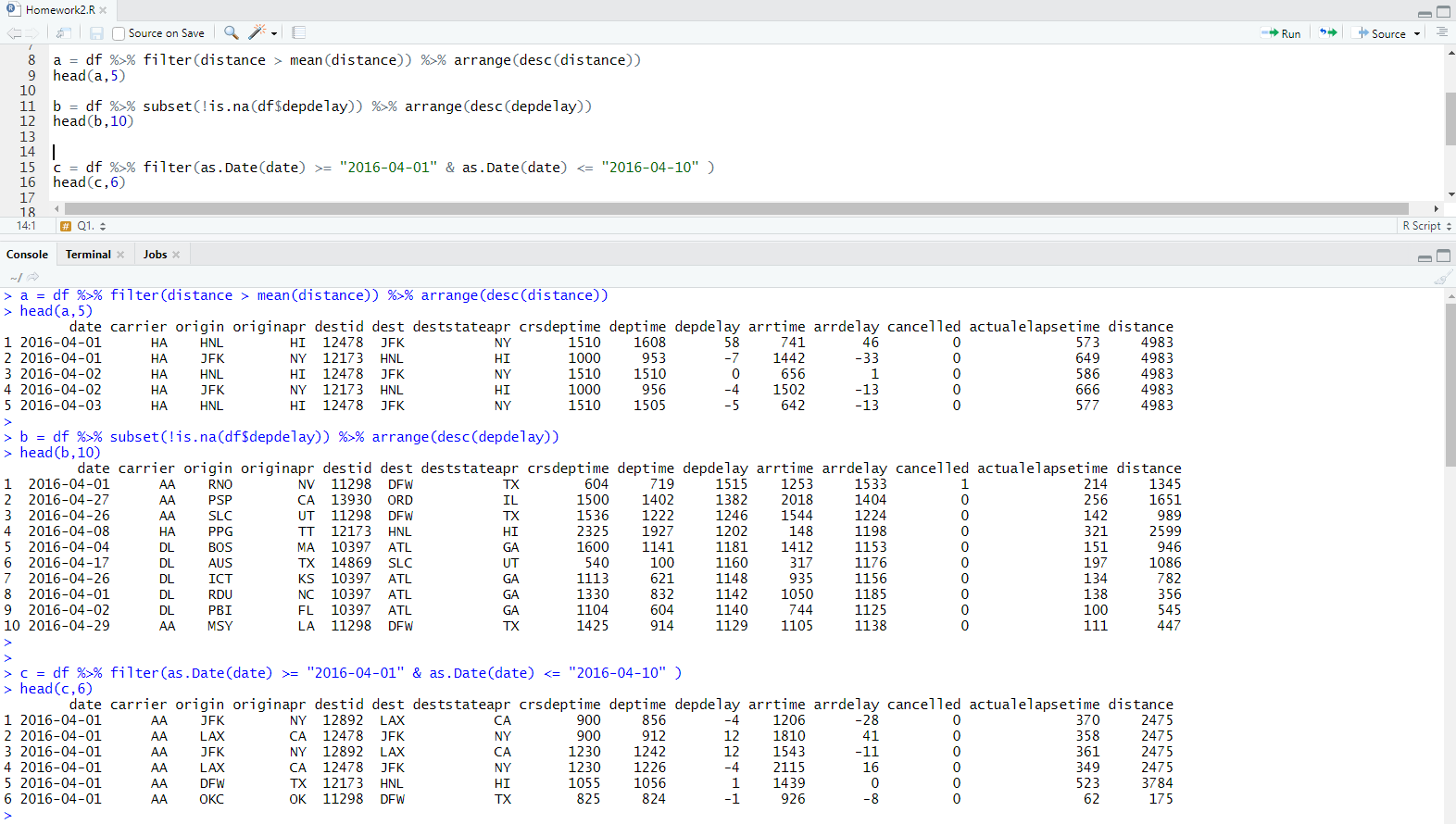
d = df %>% group\_by(carrier) %>% summarise(rate = mean(cancelled)) %>% arrange(desc(rate))

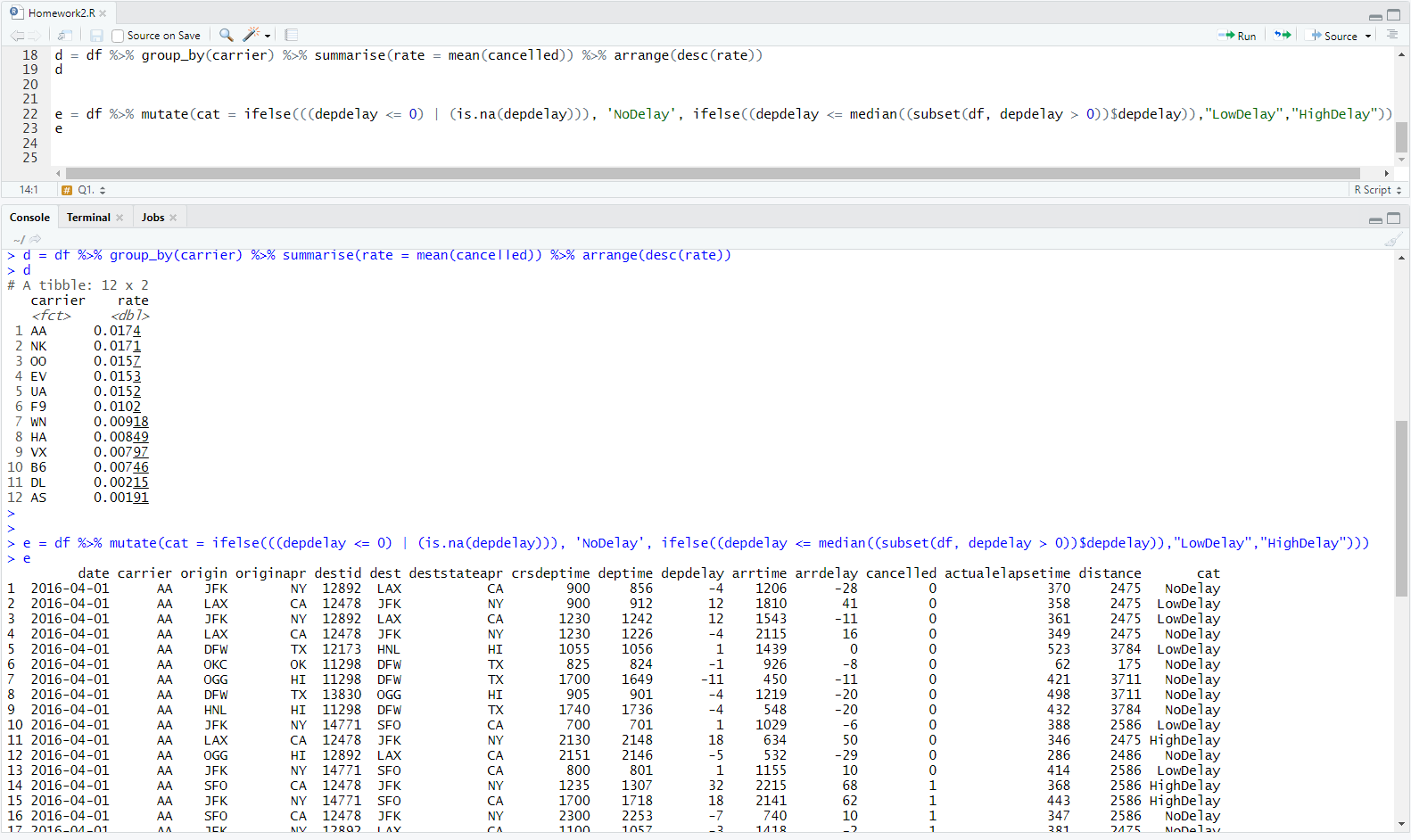
d

e = df %>% mutate(cat = ifelse(((depdelay <= 0) | (is.na(depdelay))), 'NoDelay', ifelse((depdelay <= median((subset(df, depdelay > 0))$depdelay)),"LowDelay","HighDelay")))

e

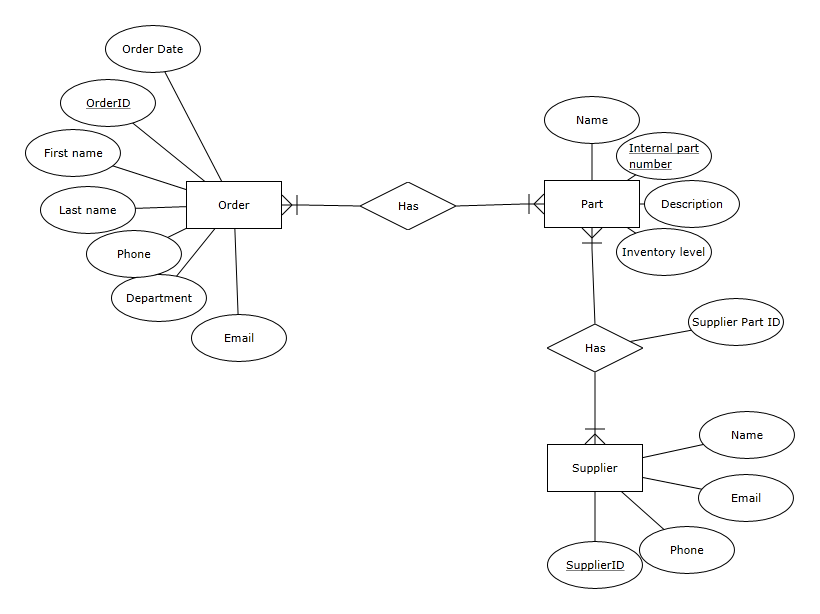
**OUTPUT**

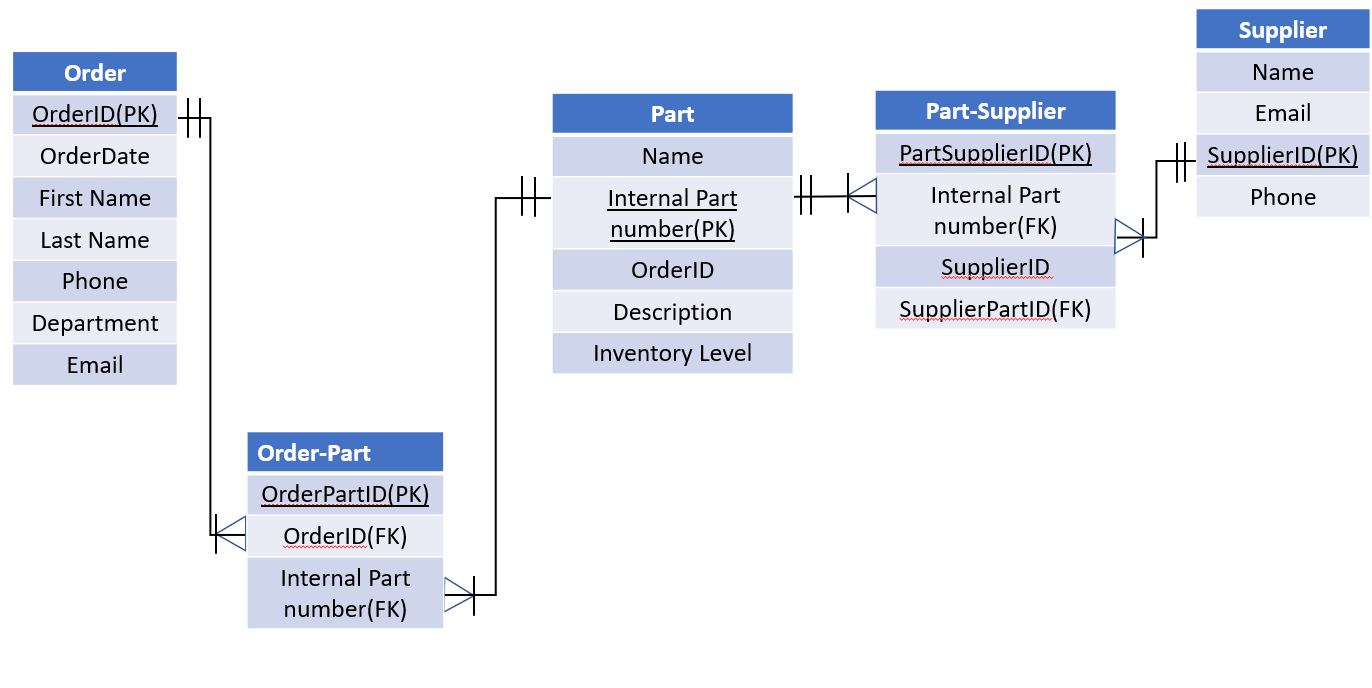




2. Create database schemas. Draw database schemas for the following scenarios. You can use PowerPoint to draw schemes and take screenshots. Please do not forget to specify **primary keys** and **foreign keys**, as well as **cardinality**. (20 Points)

1. First scenario: inventory and parts. The attribute with underscore “\_” is the primary key.

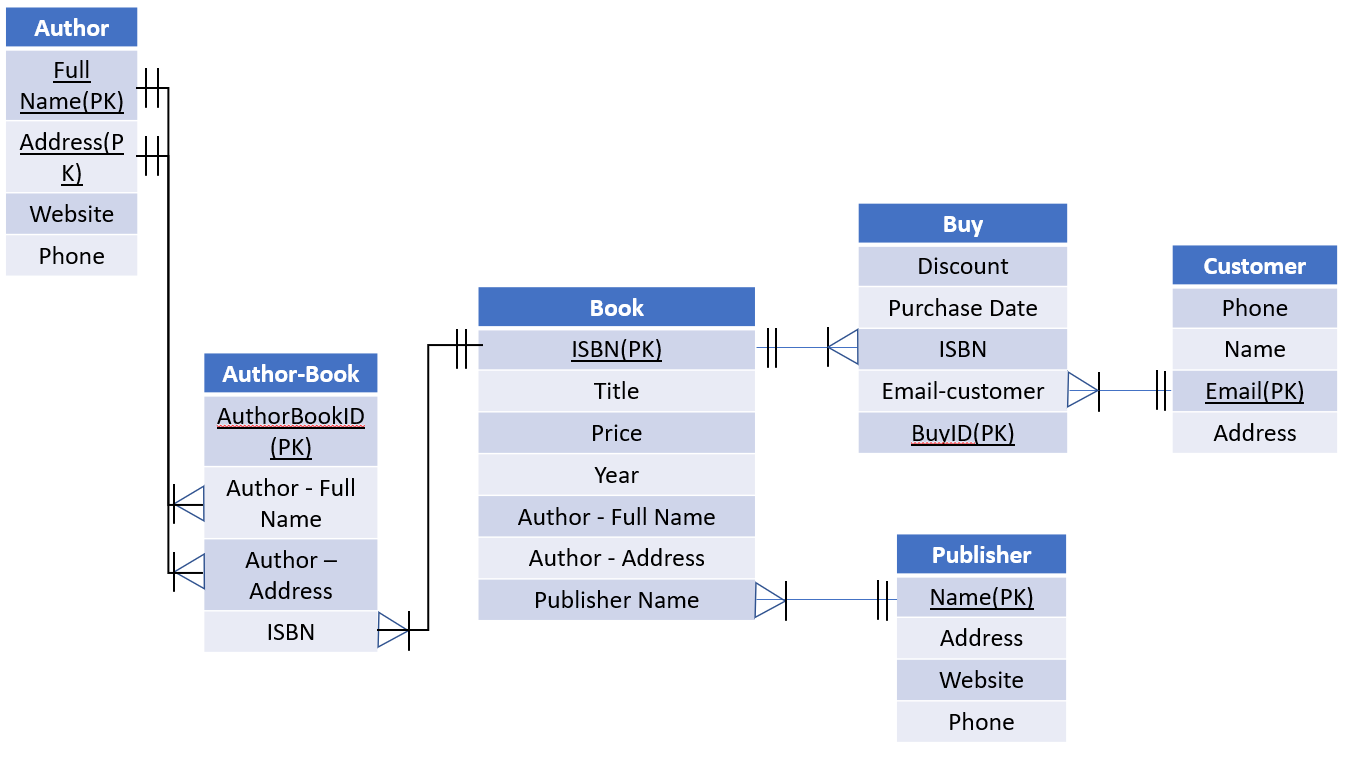




1. Second scenario: Online bookstore. The attribute with underscore “\_” is the primary key.

A close up of text on a white background

Description automatically generated



3. SQL Query. Assuming there is a database with following data schema, please use SQL queries to answer the questions below. Make sure that you only return the information needed to answer the questions (Don’t use select \*), and **only join tables if necessary**. Suppose the date format is YYYY-MM-DD and let’s assume that **date can be sorted**. (20 Points)

Scenario: hospital

A screenshot of a cell phone

Description automatically generated

1. What is the address of the doctor whose name is “Meredith Grey?” Display the name and address.

SELECT name, address FROM doctor WHERE name = “Meredith Grey”;

1. Which doctor has the largest number of prescriptions in January 2018? Display the doctor’s name and number of prescriptions. Suppose there is no tie.

SELECT doctor.name, count(prescription.prescriptionID) AS prescripCNT FROM prescription   
INNER JOIN doctor ON prescription.doctorid = doctor.doctorid  
WHERE prescription.date BETWEEN '2018-01-01' AND '2018-01-31'

GROUP BY doctor.doctorid

ORDER BY prescripCNT desc

LIMIT 1;

1. What is the average quantity of each drug in the prescriptions? Display the drug trade name and the average quantity.

SELECT tradename, AVG(quantity) as avg\_drug\_quantity FROM prescription GROUP BY tradename;

1. Which drugs are the most expensive sold by the pharmacies in Lansing? Suppose there are multiple drug-pharmacy pairs with the same price. Display the pharmacy ID, drug trade name, and price.

SLECT drugpharmacy.tradename, drugpharmacy.pharmacyID, drugpharmacy.price FROM drugpharmacy

INNER JOIN pharmacy ON drugpharmacy.pharmacyid = pharmacy.pharmacyid

Where pharmacy.city = “Lansing” AND

Drugpharmacy.price = (SELECT MAX(price) FROM drugpharmacy INNER JOIN pharmacy ON drugpharmacy.pharmacyid = pharmacy.pharmacyid WHERE pharmacy.city = “Lansing”);

(5) How many prescriptions with drug Aspirin (trade name) have the youngest female patient received? Suppose there are multiple female patients of the same age. Display patient name, gender, age, and number of prescriptions.

SELECT Patient.Name, Patient.Gender, Patient.Age, count(Prescription.PrescriptionID) as NoofPrescription from Prescription INNER JOIN Patient ON Prescription.PatientID = Patient.PatientID

WHERE Prescription.TradeName = 'Aspirin' and Patient.Gender = 'Female'

and Patient.age = (SELECT MIN(Patient.age) from Prescription

INNER JOIN Patient b on Prescription.PatientID = Patient.PatientID

WHERE Prescription.TradeName = 'Aspirin' AND Patient.Gender = 'Female'

)

GROUP BY Patient.Name, Patient.Gender, Patient.Age;