

School of Computer Science

Data Wrangling in Fulfilment of

DATA9910

Maksymilian Drzezdzon

C15311966

Degree: TU060/1

Module Coordinator: Brendan Tierney

Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged.

**Date: 2020/11/25**

Section A – Data Analysis

This section will focus on examining a medical insurance dataset of 1338 rows and 7 columns.[1] Collected variables include age, sex, BMI, number of children, does the person smoke, region of residence within the US and their annual insurance charges.

1. CREATE TABLE INSURANCE
2. (
3. AGE NUMBER(38,0),
4. SEX VARCHAR(50),
5. BMI NUMBER(38,3),
6. CHILDREN NUMBER(38,0),
7. SMOKER VARCHAR(50),
8. REGION VARCHAR(50) ,
9. CHARGES NUMBER(38,5)
10. );

Script to create and import table into SQL developer.

The main question for this part of the analysis is “what factors affect medical insurance costs the most”. With a limited amount of time for each section only a few areas will be explored to provide some insight meanwhile showcasing SQLs analytical functions with hypothesis testing, correlations and descriptive analytics etc.

Results for each query will be added within the code snippet for single row and column results using a /\* result here\*/ or with another snippet for multi row and column results, adding limit 1 to a query that only returns a single value is redundant.

A five-number summary is a set of descriptive statistics used to analyse a dataset; this was completed on each column of numeric data. A five-number summary consists of the min, max, median and first and third quartiles of data, a template would look like:

The next five queries are part of the five-number summary and are run against each numeric column to get a better understanding of the dataset, the column below is the charges column, representing annual medical insurance charges.

select round(max(charges),2) "max charges cost" from insurance; /\* 63,770.43 dollars \*/

The function above selects the charges from the insurance table, gets the highest value and rounds it to the 2nd decimal place and slaps on a “max charges cost” label.

select round(min(charges),2) "min charges cost" from insurance; /\* 1,121.87 dollars \*/

The function selects the charges from the insurance table, gets the lowest value and rounds it to the 2nd decimal place and adds a “min charges cost” label.

select median(charges) "middle charges cost value" from insurance; /\* 9,382.03 dollars \*/

The function above selects the charges from the insurance table, gets the median value which separates the second and third quartiles, then adds a “middle charges cost” label.

The above three queries were just used to get an idea of potential outliers.

1. select round(sum(charges), 2)"insurance costs in first quartile" from (select charges, ntile(4) over(order by charges) as QUARTILE from insurance) where quartile = 1; /\* 955,784.96 dollars \*/

2 select round(sum(charges), 2)"insurance costs in third quartile" from (select charges, ntile(4) over(order by charges) as QUARTILE from insurance) where quartile = 3; /\* 4,050,700.59 dollars \*/

The two above queries sum up and round to two decimal places any data that falls into the bottom twenty five percent. This is completed by the nested query that selects the charges column and puts it into four buckets with ntile(4) followed by selecting the charges rows with the over() function and the ordering it by charges from the insurance table. Lastly its tagged as quartile and used to grab the first and third quartile from the four buckets with where quartile is equal to three. An appropriate label is added to each resulting column.

The goal of this query was to get a more tangible idea of how much of the overall cost is under which quartile, if sum() wasn’t used the query would return a list of all records with an ntile number which isn’t interpretable.

3. select STATS\_ONE\_WAY\_ANOVA(bmi, charges, 'F\_RATIO') f\_ratio from insurance; /\*1.11\*/

/\*an f ratio close to 1 means that the null hypothesis is true, meaning there is no correlation between the two\*/

4. select STATS\_ONE\_WAY\_ANOVA(age, charges, 'SIG') p\_value from insurance; /\* 0.0000000000000001 \*/

/\* a p-value of more than 0.05 is not statistically significant meaning there is evidence for age having an impact on health insurance charges\*/

 These two queries where used to calculate the significance toward finding out what values have a statistical significance on insurance costs. However, what was a profound predictor of medical insurance was age and if the person smoked or not, both yielding a p value of close to 0. [3]

update insurance set smoker = 0 where smoker = 'no';

update insurance set smoker = 1 where smoker = 'yes';

5. select STATS\_ONE\_WAY\_ANOVA(smoker, charges, 'SIG') p\_value from insurance;

update insurance set smoker = 'no' where smoker = '0';

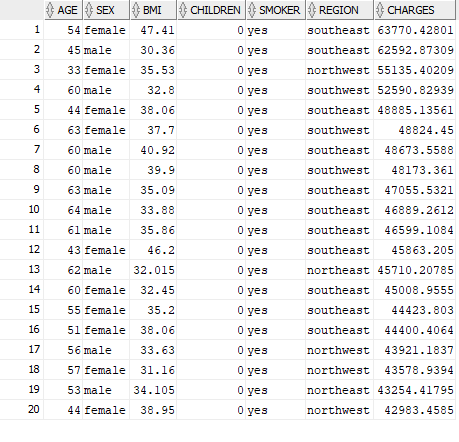
update insurance set smoker = 'yes' where smoker = '1';

 A quick test was run to make sure that the values yes and no were being interpreted properly by SQL and were converted to a 1 or 0 for the one-way anova, the result was still statistically significant

6. select STATS\_F\_TEST(smoker, charges) two\_tailed\_test from insurance; /\* 0.0000000000000000000000000000000000000000000000000003 \*/

This is the further illustrated by the two tailed test, a two tailed test ranges from 0 to 1 the closer the result is to 0 the more statistically significant it is that smoking does have an impact on medical charge costs. Alternately the ‘TWO\_SIDED\_SIG’ argument can be used which would yield the same result as a two tailed test is run by default when no test parameter is passed in. [2] The goal here was to confirm findings from the one-way ANOVA with a two tailed test.

select \* from insurance order by charges desc fetch next 20 rows only;



The top 20 highest medical insurance charges are all middle aged or older smokers, with no kids. The goal was to see if the above statistics are true and what parameters each individual would have when paying a premium on insurance.

The query selects all rows from the insurance table and orders it by the charges column limiting it to 20 rows, originally the query had children in it but as all individuals have no kids that variable was then removed as its redundant.

7. select PERCENT\_RANK(50, 'yes') within group

(ORDER BY age, smoker)

from insurance; /\*0.73\*/

7.5 select PERCENT\_RANK(50, 'no') within group

(ORDER BY age, smoker)

from insurance; /\*0.71\*/

 The goal of the above query was to see the significance of smoking on an individual ranking trying this out across a range of ages its usually 1-3%. The query takes a percent rank of an individual at the age of 50 that smokes from the group ordered by age and smoke columns from the insurance table.

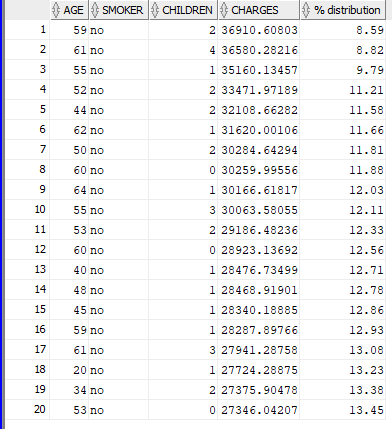
8. select age, smoker, children, charges,

round(cume\_dist() over (order by charges desc)\*100,2)"% distribution"

from insurance

order by smoker asc, "% distribution" asc

fetch next 20 rows only;



The above query grabs a cumulative distribution of individuals by their medical charges, smoking habits and % of distribution in descending order. The goal was to validate previous findings.

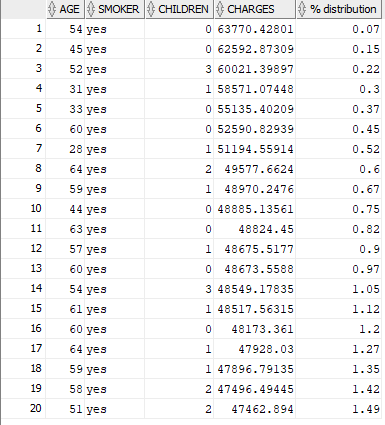
9. select age, smoker, children, charges,

round(cume\_dist() over (order by charges desc)\*100,2)"% distribution"

from insurance

order by charges desc, "% distribution" asc

fetch next 20 rows only;



When results are sorted by charges rather than whether someone is a smoker or not, the difference in medical charges almost doubles in some cases. Individuals in these groups cover a low amount of the overall however

10. select age, smoker, children, charges,

round(cume\_dist() over (order by charges desc)\*100,2)"% distribution"

from insurance

where (age = 18 and smoker = 'yes')

order by charges desc

fetch next 20 rows only;

select age, smoker, children, charges,

round(cume\_dist() over (order by charges desc)\*100,2)"% distribution"

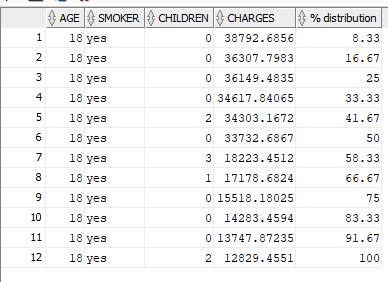
from insurance

where (age = 18 and smoker = 'no')

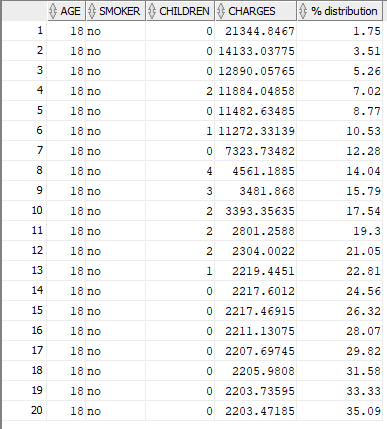
order by charges desc

fetch next 20 rows only;

when compared to eighteen-year-olds that smoke, which there are fewer categories and a much higher distribution.



When compared to the same group but instead that dont smoke, its much more spread out with lower distributions. The goal of these last few queries was to get an idea of how many people are paying more for insurance charges and have unhealthy habits, smoking and BMI were used to asses that. Maybe a campaign that offered younger customers health insurance discounts if they didn’t smoke could incentivise people to look after their health and save money. This might not be at the benefit of the insurance company but it’s an observation nonetheless.



Section B – Data Audit Report

A table was created with the following, some names had to be change because they interfered with SQL command names, such as job was changed to job\_type, day was changed to cday for campaign day etc etc.

The dataset used is from a bank telemarketing campaign that tries to sell long term deposits as a means of bouncing back from the 2008 recession.

CREATE TABLE BANK

(

age NUMBER,

job\_type VARCHAR(255),

marital VARCHAR(255),

education VARCHAR(255),

credit\_default VARCHAR(255),

balance NUMBER,

housing VARCHAR(255),

loan VARCHAR(255),

contact VARCHAR(255),

cday VARCHAR(255),

cmonth VARCHAR(255),

cduration NUMBER,

campaign NUMBER,

pdays NUMBER,

previous NUMBER,

poutcome VARCHAR(255),

y VARCHAR(255)

);

 Tasks:

Each column needs to be checked for incompleteness, empty values, is data consistent? Does it need transformations? Does it have dark data (can’t be used because it isn’t good enough).

set SERVEROUTPUT ON;

needs to be set in order for print output to be visible in the console for analysis.

This is the script that was used to asses’ data for each column, replacing age with marital, education, job\_type. The format below was used to asses values in the dataset, because most of these columns are categorical its easy to search for unique input and check for anomalies.

declare cursor cursor1

is

select distinct age from bank order by age;

record1 cursor1%rowtype;

begin

for record1 in cursor1 loop

DBMS\_OUTPUT.PUT\_line('data returned ' || record1.age);

end loop;

end;

Job categories returned admin.

Job categories returned blue-collar

Job categories returned entrepreneur

Job categories returned housemaid

Job categories returned management

Job categories returned retired

Job categories returned self-employed

Job categories returned services

Job categories returned student

Job categories returned technician

Job categories returned unemployed

Job categories returned unknown

Marital status returned divorced

Marital status returned married

Marital status returned single

Education returned primary

Education returned secondary

Education returned tertiary

Education returned unknown

Through this process data was assessed for consistency, any outliers etc. No inconsistencies, formatting errors or empty values were found, meaning that it has passed the first check. From this procedure it was found that the contact, cday, cmonth, campaign, pdays columns could be dropped and potentially reduce the storage needed as those field aren’t very useful and would qualify for being dark data. [4] Contact details are kept in order to contact customers but for the sake of training models they don’t provide any use.

If need be they can be removed with:

Alter Table bank

Drop column <col\_name\_from above>

  Alternately for data cleaning rows with an Unknown status can be deleted with

Delete from bank where education = ‘unknown’;

Delete from bank where job\_type = ‘unknown’;

   The admin occupation was cleaned up as it has a ‘.’ At the end of it for the sake of consistency with the command below.

UPDATE bank

SET job\_type = 'admin'

WHERE job\_type = 'admin.';

/\* 5,171 rows updated. \*/

declare

sig number;

mean number := 1;

stdev number := 1;

begin

SYS.dbms\_stat\_funcs.normal\_dist\_fit ('mdrzezdzon', 'bank', 'age', 'KOLMOGOROV\_SMIRNOV', mean, stdev, sig);

SYS.dbms\_stat\_funcs.POISSON\_DIST\_FIT ('mdrzezdzon', 'bank', 'age', 'KOLMOGOROV\_SMIRNOV', stdev, sig);

SYS.dbms\_stat\_funcs.normal\_dist\_fit ('mdrzezdzon', 'bank', 'balance', 'KOLMOGOROV\_SMIRNOV', mean, stdev, sig);

SYS.dbms\_stat\_funcs.POISSON\_DIST\_FIT ('mdrzezdzon', 'bank', 'balance', 'KOLMOGOROV\_SMIRNOV', stdev, sig);

SYS.dbms\_stat\_funcs.normal\_dist\_fit ('mdrzezdzon', 'bank', 'cduration', 'KOLMOGOROV\_SMIRNOV', mean, stdev, sig);

SYS.dbms\_stat\_funcs.POISSON\_DIST\_FIT ('mdrzezdzon', 'bank', 'cduration', 'KOLMOGOROV\_SMIRNOV', stdev, sig);

end;

|  |  |  |
| --- | --- | --- |
| Columns | Normal Distribution | Poisson Distribution |
| Age | 0.99 | 0.99 |
| Balance | 0.83 | 0.82 |
| Cduration | 0.99 | 0.99 |

Later columns were tested on how well they are distributed by fitting a normal and poisson distribution, this is done for the next section to make check data quality for modeling. It seems that data is normally distributed, there are no missing values, the balance column is of less importance but it was checked for anomalies.

Its worth mentioning that you can also gather performance statistics among others but that is out of scope for this project, an example is added below.

declare

begin

dbms\_stats.gather\_table\_stats('mdrzezdzon', 'bank', estimate\_percent => dbms\_stats.auto\_sample\_size);

end;

DECLARE

v\_age bank.age%type;

v\_mode\_age bank.age%type;

v\_avg\_age bank.age%type;

v\_job\_type bank.job\_type%type;

v\_marital bank.marital%type;

begin

select round(STDDEV(age),2)

into v\_age

from bank;

select round(stats\_mode(age),2)

into v\_mode\_age

from bank;

select round(AVG(age),2)

into v\_avg\_age

from bank;

select stats\_mode(job\_type)

into v\_job\_type

from bank;

select stats\_mode(marital)

into v\_marital

from bank;

DBMS\_OUTPUT.put\_line('Standard of deviation Age: ' || v\_age);

DBMS\_OUTPUT.put\_line('Average Age: ' || v\_avg\_age);

DBMS\_OUTPUT.put\_line('Most Common Age: ' || v\_mode\_age);

DBMS\_OUTPUT.put\_line('Most Common Job: ' || v\_job\_type);

DBMS\_OUTPUT.put\_line('Most Common Marital status: ' || v\_marital);

end;

Standard of deviation Age: 10.62

Average Age: 40.94

Most Common Age: 32

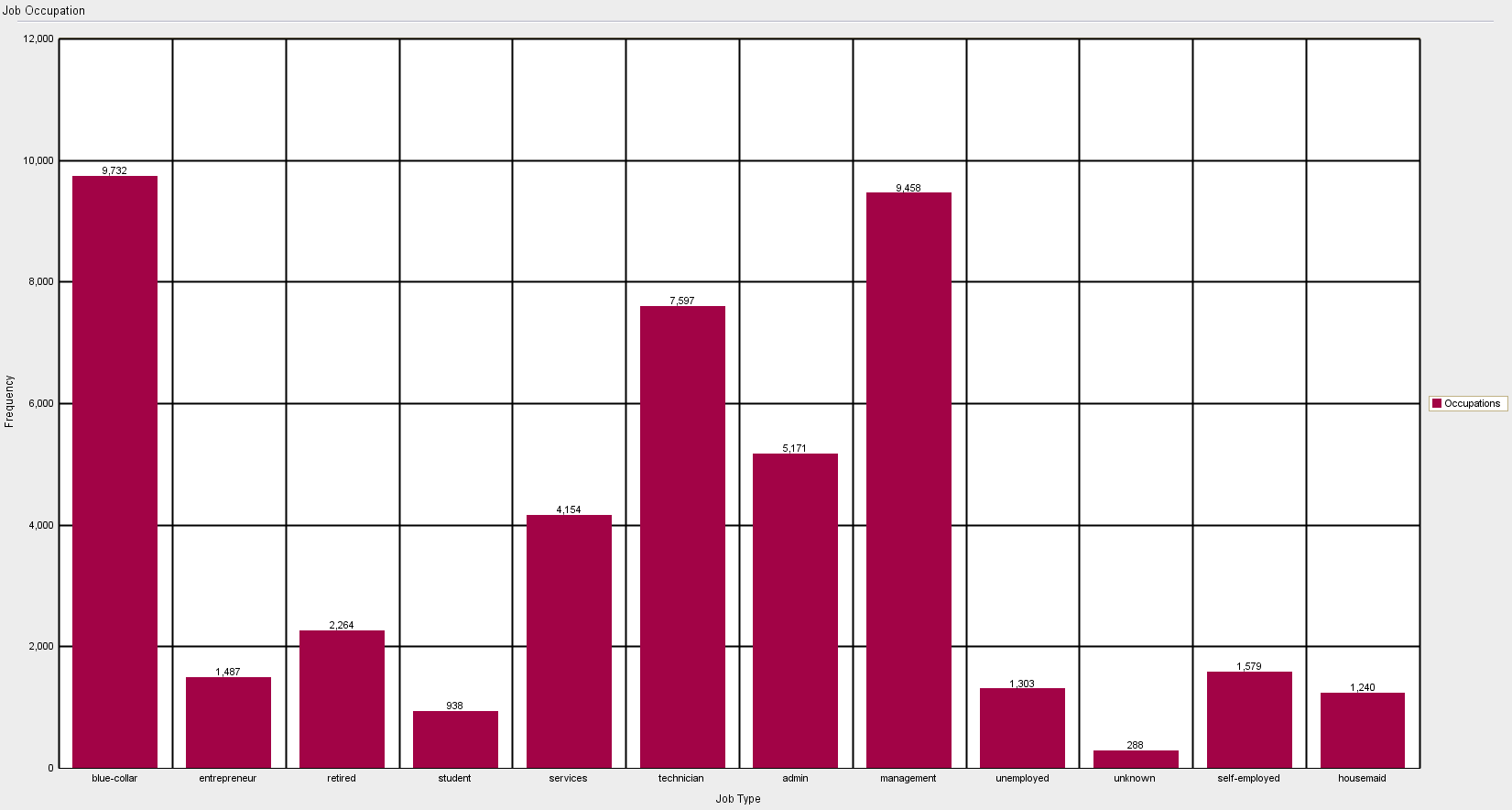
Most Common Job: blue-collar

Most Common Marital status: married

A useful tool I found while working on this report is the SQL developers report function, I won’t be focusing on it too much because I don’t think it applies to the requirements of the project but I think its worth mentioning, the bar graph was generated with SQL via SQL developer, I think it’s a better way to assess the distribution of categorical data.

Usually the 'Occupations' wouldn’t be used but in order to generate the report an additional column must be added to map the count(job\_type) result and job\_type data.

select job\_type, 'Occupations', count(job\_type) from bank group by job\_type;



In PL/SQL it would look like

set serveroutput on;

DECLARE

begin

DBMS\_OUTPUT.PUT\_LINE('');

DBMS\_OUTPUT.PUT\_LINE('Job Type Distribution');

for rec in (select job\_type, round((Count(job\_type)\* 100 / (Select Count(\*) From bank)), 2)"%" from bank group by job\_type) loop

DBMS\_OUTPUT.PUT\_LINE('| ' || rec.job\_type || ' | %' || rec."%" || '|');

end loop;

DBMS\_OUTPUT.PUT\_LINE('');

DBMS\_OUTPUT.PUT\_LINE('Marital Status Distribution');

for rec in (select marital, round((Count(marital)\* 100 / (Select Count(\*) From bank)), 2)"%" from bank group by marital) loop

DBMS\_OUTPUT.PUT\_LINE('| ' || rec.marital || ' | %' || rec."%" || '|');

end loop;

DBMS\_OUTPUT.PUT\_LINE('');

DBMS\_OUTPUT.PUT\_LINE('Education Distribution');

for rec in (select education, round((Count(education)\* 100 / (Select Count(\*) From bank)), 2)"%" from bank group by education) loop

DBMS\_OUTPUT.PUT\_LINE('| ' || rec.education || ' | %' || rec."%" || '|');

end loop;

end;

Results are below

Job Type Distribution

| blue-collar | %21.48|

| retired | %4.97|

| entrepreneur | %3.27|

| student | %1.79|

| services | %9.27|

| technician | %17.03|

| admin | %11.58|

| management | %21.34|

| unemployed | %2.95|

| self-employed | %3.57|

| housemaid | %2.77|

Marital Status Distribution

| married | %60.07|

| divorced | %11.64|

| single | %28.29|

Education Distribution

| tertiary | %30.7|

| primary | %15.74|

| secondary | %53.55|

There doesn’t seem to be anything unusual with the percentage distributions which makes it ok to use for analysis, the unknown values were removed as its data that isn’t very useful, other variables potentially could have been used to guess each unknown occupation, however these rows were few in number so they were just removed from the dataset.

Section C – Machine Learning

Appendix

Oracle 1996, 2007 <https://docs.oracle.com/database/121/SQLRF/functions186.htm#SQLRF06318> [2]

<https://docs.oracle.com/database/121/SQLRF/functions190.htm#SQLRF06322> [3]

https://www.oracletutorial.com/oracle-analytic-functions/

https://www.ibm.com/support/knowledgecenter/en/SSHRBY/com.ibm.swg.im.dashdb.apdv.plsql.doc/doc/c0053861.html

References

Miri Choi, 2017, [https://www.kaggle.com/mirichoi0218/insurance [1](https://www.kaggle.com/mirichoi0218/insurance%20%5b1)]

Christopher Tozzi, January 2020 <https://www.precisely.com/blog/data-quality/how-to-measure-data-quality-7-metrics> [4]