Assignment 3

Task 1

The following code is used to extract the difference between the left join of the company data (AS_merged) and the exploration data (expl_oper)

```
expl_AS_left_join <- expl_oper %>% left_join(AS_merged, by=c("NPD_id", "year"))

diff_set <- setdiff(expl_AS_left_join, expl_AS) %>% filter(is.na(municipality))

nrow(diff_set)
```

```
## [1] 93
```

Why were they dropped?

The diference is observed by taking the difference of a inner join of the variables NPD_id and year, and a left join of the same variables. 93 observations thus appear in the exploratory dataset that does not occur in the company dataset For a left join / merge between two rows to be made both the NPD_id and year much find an equal pairing in the other data set.

What do they have in common?

The loss of data appears random and this may limit the extent of the problems caused by the lacking of data in the company dataset.The difference appears to be largely data that is missing in the company data by happenstance. A join based on the exploratory data will thus lead to dropping of all rows where the company dataset does not match in year and NPD_id.

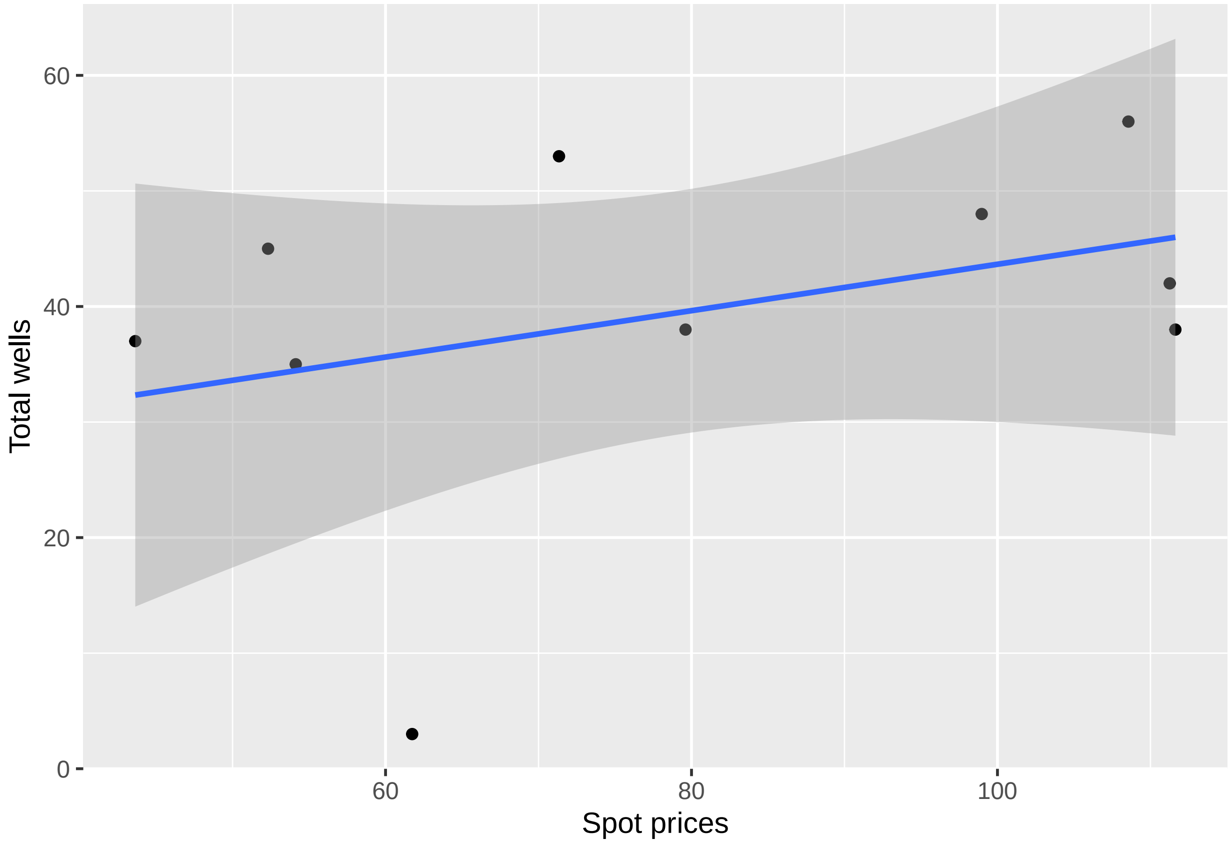
TASK 2

We plot a simple linear regression of the total exploratory wells against the spot price at the time of the wells completion.

```
brent_spotprices <- read.csv("Europe_Brent_Spot_Price_FOB.csv") %>% tail(., -4)

brent_spotprices %<>%
  mutate(year = as.numeric(rownames(brent_spotprices)),
         Europe.Brent.Spot.Price.FOB = as.numeric(Europe.Brent.Spot.Price.FOB)) %>%
  rename(spot_price = Europe.Brent.Spot.Price.FOB)


explAGG_spot <- explAgg_AS %>% inner_join(brent_spotprices, by = c("year"))
```

```
reg_spot_data <- explAGG_spot %>% group_by(spot_price) %>%
  summarise(total_n_wells = sum(numbWells))

reg_spot_data %>%
  ggplot(aes(x= spot_price, y= total_n_wells)) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("Spot prices") +
  ylab("Total wells")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



We conduct a simple linear regression.

```
spotprice_regression_total_wells <- lm(formula = total_n_wells ~ spot_price, data = reg_spot_data)
summary(spotprice_regression_total_wells)
```

```
##
## Call:
## lm(formula = total_n_wells ~ spot_price, data = reg_spot_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.969  -3.326   2.556   9.137  15.103
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.5668    15.1361   1.557    0.158
## spot_price    0.2009     0.1819   1.104    0.302
##
## Residual standard error: 14.46 on 8 degrees of freedom
## Multiple R-squared:  0.1323, Adjusted R-squared:  0.02379
## F-statistic: 1.219 on 1 and 8 DF,  p-value: 0.3016
```

There seems to be a small, but insignificant correlation between the spot price of brent oil and the total number of exploratory wells. As we can see from the p-values, it is highly likely to be coincidental. It should be said that there are quite few data points, so each deviation have a big impact on the result.

TASK 3

Overview of the CO2 taxes in the period 2009-2020.

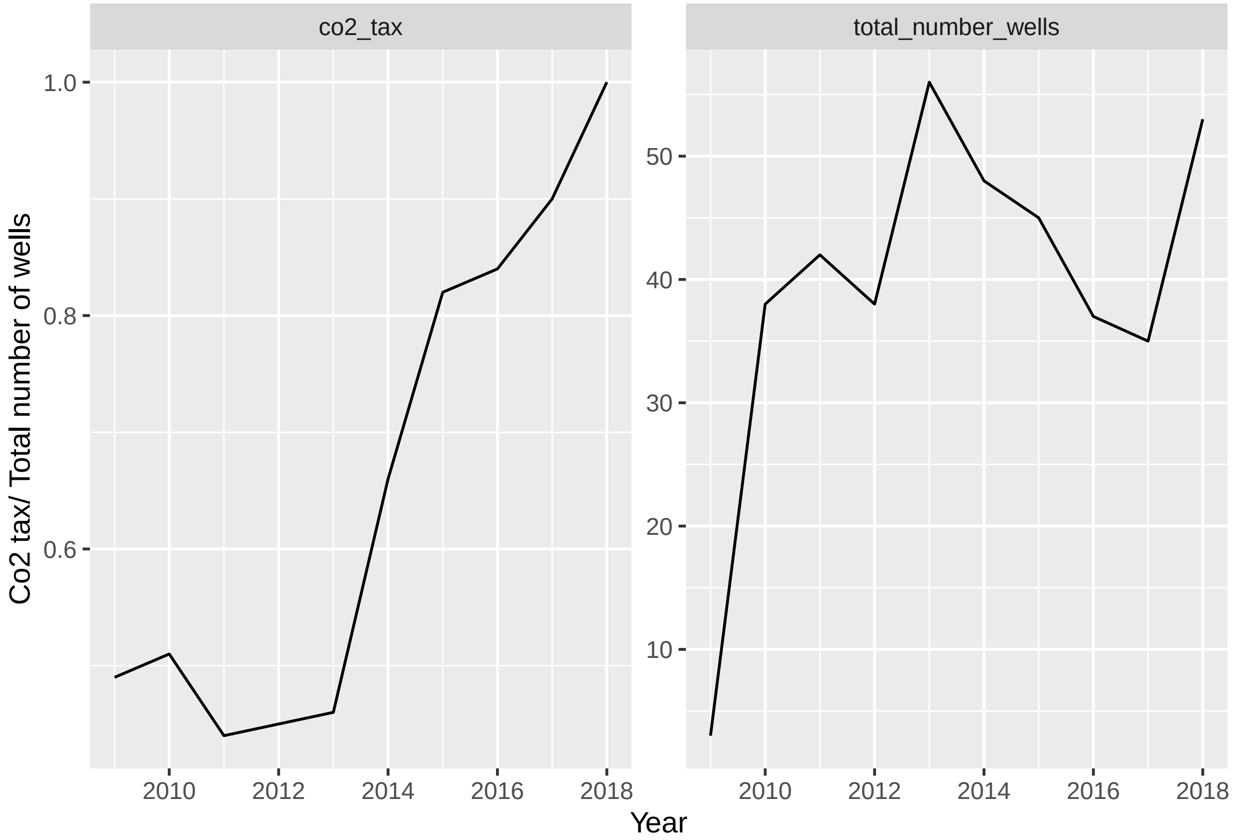
```
CO2_tax = tibble(
  year=2009:2020,
  CO2_tax_NG_krSM3=c(0.49, 0.51, 0.44, 0.45, 0.46, 0.66,.82, 0.84, 0.90, 1, 1, 1.08)
  )
```

The exploratory drilling seems to be unaffected by the hike in the Co2 taxes. Simply plotting the number of exploratory wells against the the increase Co2 tax levels, does not seem to correspond with the rate or time period in which the number of exploratory wells increase.

```
## Run regression tests

num_wells <- inner_join(explAgg_AS, CO2_tax, by = "year")  %>%
  group_by(year) %>%
  summarise(total_number_wells = sum(numbWells),
            co2_tax = CO2_tax_NG_krSM3) %>%
  unique()
```

```
## ‘summarise()‘ has grouped output by ’year’. You can override using the ‘.groups‘ argument.
```

```
num_reg <- lm(total_number_wells ~ co2_tax, data = num_wells)


## Plot CO2 tax vs exploration

num_wells_pivot <- num_wells %>% pivot_longer(cols = c(co2_tax, total_number_wells),
                                              names_to = "variable",
                                              values_to = "variable_values")

num_wells_pivot %>%
  ggplot(aes(x = year, y = variable_values)) +
  geom_line() +
  facet_wrap(~ variable, scales = "free_y") +
  xlab("Year") +
  ylab("Co2 tax/ Total number of wells")
```



Conducting a simple linear regression tests suggest that the is no statistically significant correlation between tax rate and the total number of exploratory wells.

```
summary(num_reg)
```

```
##
## Call:
## lm(formula = total_number_wells ~ co2_tax, data = num_wells)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.750  -3.905   2.362   7.407  19.744
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    28.68      16.08   1.784    0.112
## co2_tax        16.47      23.37   0.705    0.501
##
## Residual standard error: 15.06 on 8 degrees of freedom
## Multiple R-squared:  0.05843,    Adjusted R-squared:  -0.05926
## F-statistic: 0.4965 on 1 and 8 DF,  p-value: 0.5011
```

However, this lack of correlation may simply be due to time inconsistencies, e.g exploratory drilling requires a lot of planning and a tax hike will only affect future drilling and not the total number of wells in the near future.

Task 4

The following regression is conducted.

```
reg1 = lm(numbWells ~ total_assets + profitability, data=explAgg_AS)
summary(reg1)
```

```
##
## Call:
## lm(formula = numbWells ~ total_assets + profitability, data = explAgg_AS)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8817 -1.9410 -1.3936  0.4946 11.2441
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.428e+00  3.798e-01   6.391 5.13e-09 ***
## total_assets   2.673e-08  3.491e-09   7.656 1.18e-11 ***
## profitability -4.859e-03  1.015e-02  -0.479    0.633
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.47 on 101 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.3789, Adjusted R-squared:  0.3667
## F-statistic: 30.81 on 2 and 101 DF,  p-value: 3.572e-11
```

A

The linear regression is composed of two predictor variables: asset size and the profitability of the companies. From the regression printout we can see that the more assets a company has the more exploratory drilling we can expect the company to conduct. The coefficient of total assets can be interpreted as an expectation of nearly three more wells per 100 million NOK kroner. The profitability has a negative coefficient, but this does not imply a negative correlation as it its explanatory power is not significant as determined by the p-value.

The R-squared is quite low, this could be improved by adding more predictor variables.

B

The total assets are supplied in different currencies, NOK and USD. The interpretation of the coefficient as well as accuracy of the t-test will be impaired as long as the data is not standardized to one currency. We could not find any documentation as how profitability is measured (e.g in thousands/millions etc). This affects the interpretation of the profitability predictor.

C

We solve the currency issue by converting all relevant data of our chosen variables to NOK, and conduct a new regression.

```
explAgg_AS_single_currency <- explAgg_AS %>%
  mutate(
         total_assets = case_when(currency_code != "NOK" ~ as.double(total_assets * 8.52),
                                  TRUE ~ as.double(total_assets)),
         total_debt   = case_when(currency_code != "NOK" ~ as.double(total_debt * 8.52),
                                  TRUE ~ as.double(total_assets)),
         total_operating_costs   = case_when(currency_code != "NOK" ~ as.double(total_operating_costs *
                                  TRUE ~ as.double(total_operating_costs)),
         sales_revenue   = case_when(currency_code != "NOK" ~ as.double(sales_revenue * 8.52),
                                  TRUE ~ as.double(sales_revenue)),
         currency_code = "NOK")


reg2 = lm(numbWells ~ total_debt  + sales_revenue  , data= explAgg_AS_single_currency)
summary(reg2)
```

```
##
## Call:
## lm(formula = numbWells ~ total_debt + sales_revenue, data = explAgg_AS_single_currency)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4955 -1.4841 -0.9927  0.8089 11.5877
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.954e+00  3.563e-01   5.485 3.02e-07 ***
## total_debt     4.086e-08  7.897e-09   5.175 1.15e-06 ***
## sales_revenue -3.010e-08  1.523e-08  -1.976   0.0509 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.168 on 102 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.4808, Adjusted R-squared:  0.4706
## F-statistic: 47.22 on 2 and 102 DF,  p-value: 3.04e-15
```

We exclude total assets from the regression after discovering that its explanatory power is derived in large part from its correlation with total debt. As such total debt explains more of the number of wells expected than total assets of a company. We include the companies sales revenue in our regression model and find it is negative correlation with the number of exploratory wells. This finding is surprising.