Domas Budrys A2 – CSCI5040

Question 1:

```
val dataW = spark.read.option("inferSchema", "true").option("header",
"true").csv("/Users/domo/Desktop/spark-data/walmart.csv")

import org.apache.spark.sql.functions._

dataW.select(max("Weekly_Sales")).show(1)
```

```
+-----------------+
|max(Weekly_Sales)|
+-----------------+
|        693099.36|
+-----------------+
```

Question 2:

```
dataW.groupBy("Store","Dept").max("Weekly_Sales").withColumnRenamed("m
ax(Weekly_Sales)","Highest_Sale").sort(desc("Highest_Sale")).limit(1).
show()
```

```
+-----+----+------------+
|Store|Dept|Highest_Sale|
+-----+----+------------+
|   10|  72|   693099.36|
+-----+----+------------+
```

Question 3:

```
val dataQ3 =
dataW.groupBy("Store").avg("Weekly_Sales").withColumnRenamed("avg(Week
ly_Sales)", "Average_Sale").sort(asc("Store"))

dataQ3.withColumn("Average_Sale",
format_number(dataQ3("Average_Sale"), 2)).show()
```

```
+-----+------------+
|Store|Average_Sale|
+-----+------------+
|    1|   21,710.54|
|    2|   26,898.07|
|    3|    6,373.03|
|    4|   29,161.21|
|    5|    5,053.42|
|    6|   21,913.24|
|    7|    8,358.77|
|    8|   13,133.01|
|    9|    8,772.89|
|   10|   26,332.30|
|   11|   19,276.76|
|   12|   14,867.31|
|   13|   27,355.14|
```

Question 4:

```
val dataQ4 = dataW.selectExpr("Store","Dept", "Date",
"Weekly_Sales").where("Weekly_Sales < 0")

dataQ4.count()
```

Result → res2: Long = 1285

Question 5:

```
import java.sql.Date
import org.apache.spark.sql.types.{DateType, IntegerType}

dataW.withColumn("Year_Value",year(col("Date"))).where("Year_Value=201
1").count()
```

Result → res54: Long = 153453

Question 6:

```
dataW.selectExpr("Store","Dept", "Date", "Weekly_Sales",
"IsHoliday").withColumn("Weekly_Sales",
format_number(dataW("Weekly_Sales"),2)).show()
```

Store	Dept	Date	Weekly_Sales	IsHoliday
1	1	2010-02-05 00:00:00	24,924.50	false
1	1	2010-02-12 00:00:00	46,039.49	true
1	1	2010-02-19 00:00:00	41,595.55	false
1	1	2010-02-26 00:00:00	19,403.54	false
1	1	2010-03-05 00:00:00	21,827.90	false
1	1	2010-03-12 00:00:00	21,043.39	false
1	1	2010-03-19 00:00:00	22,136.64	false
1	1	2010-03-26 00:00:00	26,229.21	false
1	1	2010-04-02 00:00:00	57,258.43	false

Question 7:

```
val dataQ7 = dataW.selectExpr("Store","Dept", "Date", "Weekly_Sales",
"IsHoliday").where("Weekly_Sales < 0")

dataQ7.withColumn("Weekly_Sales",
regexp_replace(dataQ7("Weekly_Sales"),"-", "")).show()
```

```
+-----+----+-------------------+------------+---------+
|Store|Dept|               Date|Weekly_Sales|IsHoliday|
+-----+----+-------------------+------------+---------+
|    1|   6|2012-08-10 00:00:00|      139.65|    false|
|    1|  18|2012-05-04 00:00:00|        1.27|    false|
|    1|  47|2010-02-19 00:00:00|       863.0|    false|
|    1|  47|2010-03-12 00:00:00|       698.0|    false|
|    1|  47|2010-10-08 00:00:00|        58.0|    false|
|    1|  47|2011-04-08 00:00:00|       298.0|    false|
|    1|  47|2011-07-08 00:00:00|       198.0|    false|
|    1|  47|2011-10-14 00:00:00|       498.0|    false|
|    1|  47|2011-12-23 00:00:00|       498.0|    false|
|    1|  47|2012-02-17 00:00:00|       198.0|    false|
|    1|  47|2012-03-16 00:00:00|       199.0|    false|
|    1|  48|2012-03-23 00:00:00|       223.0|    false|
|    1|  54|2011-01-21 00:00:00|        50.0|    false|
|    1|  54|2011-05-20 00:00:00|        15.0|    false|
|    1|  54|2012-03-09 00:00:00|        21.0|    false|
|    2|  18|2012-06-01 00:00:00|        1.97|    false|
|    2|  18|2012-07-27 00:00:00|        3.03|    false|
|    2|  45|2010-04-09 00:00:00|       118.0|    false|
|    2|  45|2010-05-07 00:00:00|        0.98|    false|
|    2|  47|2010-07-30 00:00:00|      1098.0|    false|
+-----+----+-------------------+------------+---------+
only showing top 20 rows
```

Question 8:

```
val Q9DownHolidays = dataW.selectExpr("Store","Dept", "Date",
"Weekly_Sales", "IsHoliday").where("Weekly_Sales < 0 AND  isHoliday ==
true").count()
```
Result → Q9DownHolidays: Long = 98

```
val Q9AllHolidays = dataW.selectExpr("Store","Dept", "Date",
"Weekly_Sales", "IsHoliday").where("isHoliday == true").count()
```
Result → Q9AllHolidays: Long = 29661

```
val resultHolidays = Q9DownHolidays.toFloat / Q9AllHolidays.toFloat
```
Result → resultHolidays: Float = 0.0033040019

```
val Q9DownNon = dataW.selectExpr("Store","Dept", "Date",
"Weekly_Sales", "IsHoliday").where("Weekly_Sales < 0 AND  isHoliday ==
false").count()
```
Result → Q9DownNon: Long = 1187

```
val Q9AllNon = dataW.selectExpr("Store","Dept", "Date",
"Weekly_Sales", "IsHoliday").where("isHoliday == false").count()
```
Result → Q9AllNon: Long = 391909

```
val resultNoneHolidays = Q9DownNon.toFloat / Q9AllNon.toFloat
```
Result → resultNoneHolidays: Float = 0.0030287644

Based on given numbers, only 0.33% of weeks experience lost in sales during holidays. On regular week (none holidays) Walmart experienced 0.302% loss in sales