Task 2: Chi-Square hypothesis test

Objective: Based on employee sample data check if salary depends on role.

```
#Read data
df = spark.read.csv("/home/s_kante/spark/data/developers_survey_training.csv", header='true')

#Replace IsDeveloper value with integer 1 or 0
df.createOrReplaceTempView("inputData")
df1 = spark.sql("SELECT CASE IsDeveloper WHEN 'Yes' THEN 1 ELSE 0 END AS
IsDeveloper, CASE WHEN YearsOfExp<=2 THEN 1 WHEN YearsOfExp>2 AND
YearsOfExp<=5 THEN 2 ELSE 3 END AS YearsOfExp, CASE WHEN Salary<=50000 THEN 1
WHEN  Salary>50000 AND Salary<100000 THEN 2 ELSE 3 END AS Salary FROM inputData
");

#Create feature vector
from pyspark.ml.feature import VectorAssembler
assembler = VectorAssembler(inputCols=["IsDeveloper","YearsOfExp"], outputCol="features")
combined = assembler.transform(df1)
vector_df = combined.select(combined.Salary, combined.features)

#Find the chi-square stats
from pyspark.ml.stat import ChiSquareTest
chiResult = ChiSquareTest.test(vector_df, "features", "Salary")

#Display the stats
chiResult.head().pValues
chiResult.head().degreesOfFreedom
```

Ref:
https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests/chi-square-goodness-of-fit-tests/v/chi-square-distribution-introduction
https://www.ling.upenn.edu/~clight/chisquared.htm
https://stattrek.com/chi-square-test/goodness-of-fit.aspx
Best: https://stattrek.com/chi-square-test/independence.aspx