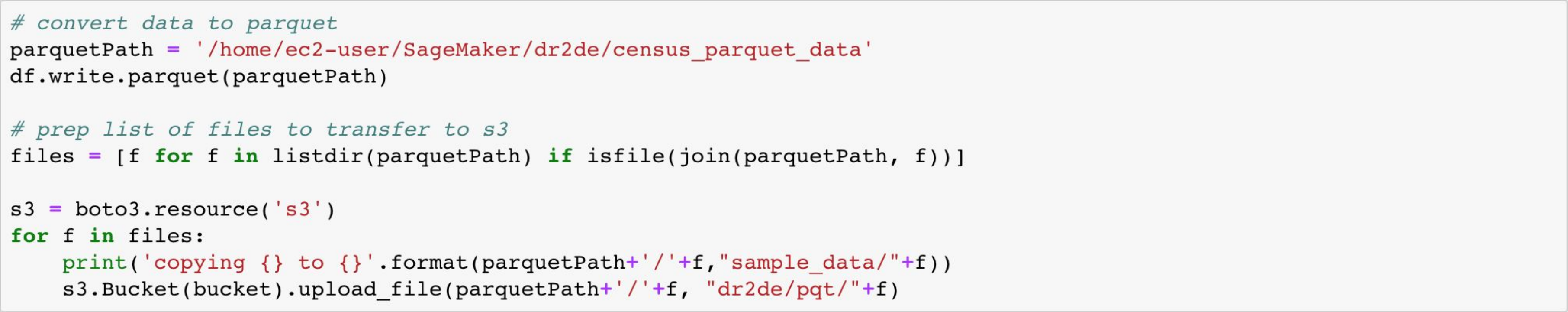
Motivation

- Census data: (https://www.kaggle.com/muonneutrino/us-census-demographic-data#acs2015_county_data.csv)
- Contains a number of demographic features pertaining to counties across the United States
- Target variable is unemployment rate

Code Snippet



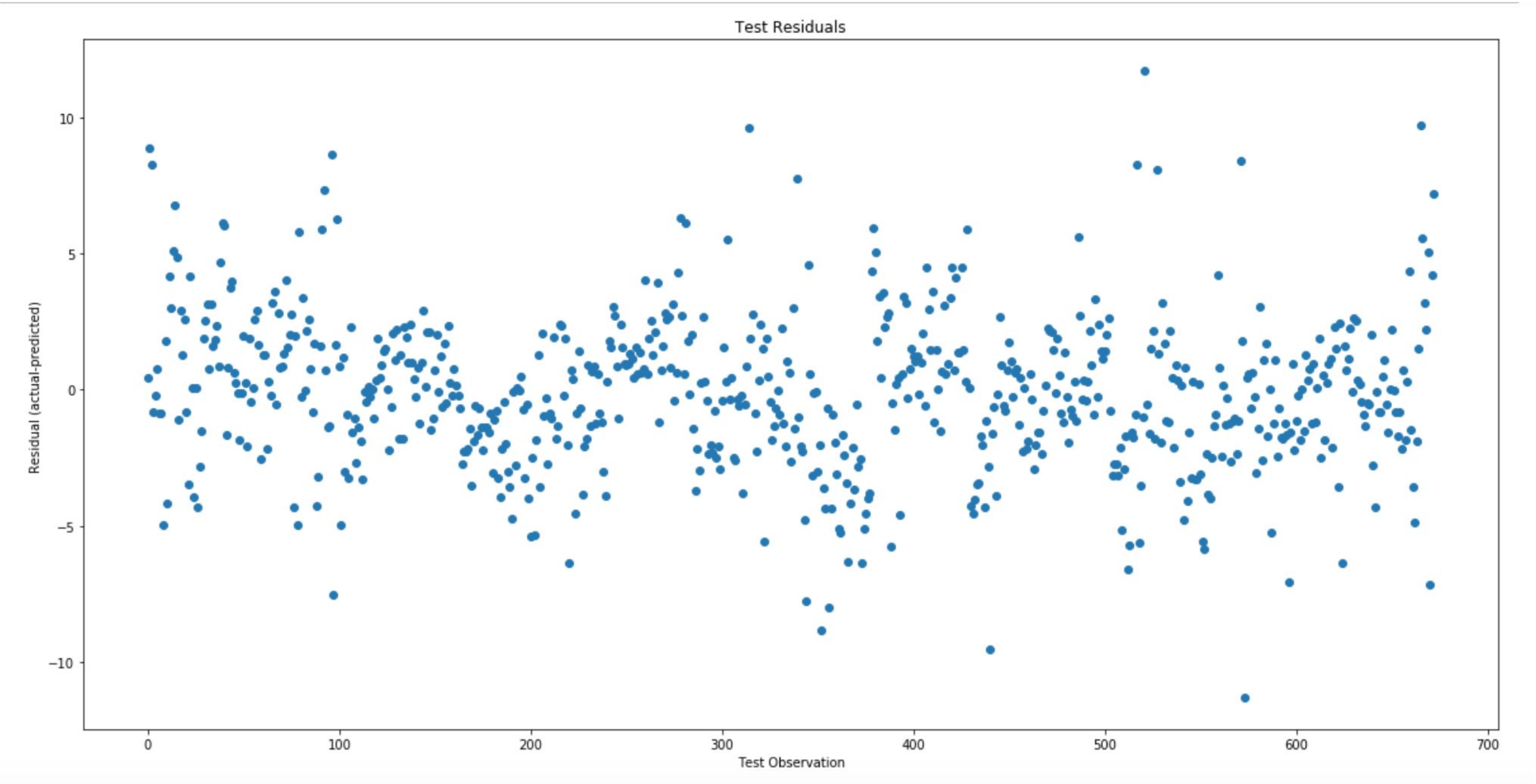
```
# convert data to parquet
parquetPath = '/home/ec2-user/SageMaker/dr2de/census_parquet_data'
df.write.parquet(parquetPath)

# prep list of files to transfer to s3
files = [f for f in listdir(parquetPath) if isfile(join(parquetPath, f))]

s3 = boto3.resource('s3')
for f in files:
    print('copying {} to {}'.format(parquetPath+'/'+f,"sample_data/"+f))
    s3.Bucket(bucket).upload_file(parquetPath+'/'+f, "dr2de/pqt/"+f)
```

- Conversion of data to parquet for efficient storage and transfer to Amazon S3

Residuals from simple linear regression (unemployment rate ~ poverty rate)


Test Residuals
Residual (actual-predicted)
Test Observation