
H2O.ai

# Group By Aggregation

```
grouped_data = census_data[["occupation", "education-num"]].group_by(["occupation"])
stats = grouped_data.count(na = "ignore").median(na = "ignore").mean(na = "ignore").sd(na = "ignore")
stats.get_frame()
```

| occupation | nrow | median_education-num | mean_education-num | sdev_education-num |
|---|---|---|---|---|
| | 0 | 9 | 9.25339 | 2.60279 |
| Adm-clerical | 3770 | 10 | 10.1135 | 1.69805 |
| Armed-Forces | 9 | 9 | 10.1111 | 2.02759 |
| Craft-repair | 4099 | 9 | 9.11076 | 2.03865 |
| Exec-managerial | 4066 | 12 | 11.4491 | 2.14321 |
| Farming-fishing | 994 | 9 | 8.60865 | 2.75607 |
| Handlers-cleaners | 1370 | 9 | 8.51022 | 2.20338 |
| Machine-op-inspct | 2002 | 9 | 8.48751 | 2.28528 |
| Other-service | 3295 | 9 | 8.77967 | 2.29966 |
| Priv-house-serv | 149 | 9 | 7.36242 | 3.11104 |

# Cross-Validated Mean Target Encoding

(Feature Engineering Sneak Peak)

```
def mean_target_encoding(data, x, y, fold_column):
    grouped_data = data[[x, fold_column, y]].group_by([x, fold_column])
    grouped_data.sum(na = "ignore").count(na = "ignore")
    df = grouped_data.get_frame().as_data_frame()
    df_list = []
    nfold = int(data[fold_column].max()) + 1
    for j in range(0, nfold):
        te_x = "te_{}".format(x)
        sum_y = "sum_{}".format(y)
        oof = df.loc[df[fold_column] != j, [x, sum_y, "nrow"]]
        stats = oof.groupby([x]).sum()
        stats[x] = stats.index
        stats[fold_column] = j
        stats[te_x] = stats[sum_y] / stats["nrow"]
        df_list.append(stats[[x, fold_column, te_x]])
    return h2o.H2OFrame(pd.concat(df_list))
```

# Group By Aggregation

```
1 grouped_data = census_data[["occupation", "education-num"]].group_by(["occupation"])
2 stats = grouped_data.count(na = "ignore").median(na = "ignore").mean(na = "ignore").sd(na = "ignore")
3 stats.get_frame()
```

| occupation | nrow | median_education-num | mean_education-num | sdev_education-num |
|---|---|---|---|---|
| | 0 | 9 | 9.25339 | 2.60279 |
| Adm-clerical | 3770 | 10 | 10.1135 | 1.69805 |
| Armed-Forces | 9 | 9 | 10.1111 | 2.02759 |
| Craft-repair | 4099 | 9 | 9.11076 | 2.03865 |
| Exec-managerial | 4066 | 12 | 11.4491 | 2.14321 |
| Farming-fishing | 994 | 9 | 8.60865 | 2.75607 |
| Handlers-cleaners | 1370 | 9 | 8.51022 | 2.20338 |
| Machine-op-inspct | 2002 | 9 | 8.48751 | 2.28528 |
| Other-service | 3295 | 9 | 8.77967 | 2.29966 |