In [3]:

```
import pandas as pd
import scipy.stats as stats
import statsmodels.api as sm
from statsmodels.formula.api import ols
import scikit_posthocs as sp
```

In [4]:

```
data_anova = pd.read_excel('/Users/yashgupta/Downloads/BSAN Project Files Edited/Final Datasets/Neonatal ANOVA.xlsx')
```

In [5]:

```
data_anova.isna().sum()
```

Out[5]:

```
Unnamed: 0                    0
HOSP_KID                      0
RECNUM                        0
HOSP_LOCTEACH                 0
H_CONTRL                      0
HOSP_REGION                   0
HOSP_BEDSIZE                  0
AMONTH                        0
AWEEKEND                      0
DQTR                          0
DQTR_DES                      0
ELECTIVE                      0
ELECTIVE_DES                  0
APRDRG                        0
APRDRG_FULL                   0
APRDRG_Risk_Mortality         0
APRDRG_Risk_Mortality_FULL    0
APRDRG_Severity               0
APRDRG_Severity_FULL          0
AGE                           0
AGE_NEONATE                   0
DISPUNIFORM                   0
DISPUNIFORM_DES               0
DIED                          0
FEMALE                        0
FEMALE_DES                    0
HOSP_REGION_DES               0
PAY1                          0
PAY1_DES                      0
RACE                          0
RACE_DES                      0
TOTCHG                        0
ZIPINC_QRTL                   0
Median Income                 0
LOS                           0
PCLASS_ORPROC                 0
PL_NCHS                       0
PL_NCHS_DES                   0
HOSP_LOCTEACH_FULL            0
H_CONTRL_FULL                 0
dtype: int64
```

ANOVA

In [6]:

```
# Check for non-numeric values in the 'TOTCHG' column
non_numeric = data_anova['TOTCHG'].apply(lambda x: isinstance(x, str))
missing_data = data_anova['TOTCHG'].isnull()

# Remove rows with non-numeric 'TOTCHG' values
clean_data_anova = data_anova[~(non_numeric | missing_data)].copy()

# Convert 'TOTCHG' to numeric
clean_data_anova['TOTCHG'] = pd.to_numeric(clean_data_anova['TOTCHG'])

# Perform the Two-Way ANOVA
model_anova_clean = ols('TOTCHG ~ C(HOSP_REGION_DES) + C(RACE_DES) + C(HOSP_REGION_DES):C(RACE_DES)', data=clean_data_anova).fit()
anova_results_clean = sm.stats.anova_lm(model_anova_clean, typ=2)

# Output the results
print(anova_results_clean)
```

```
                                      sum_sq      df         F    PR(>F)
C(HOSP_REGION_DES)              1.355665e+13     3.0  8.594334  0.000011
C(RACE_DES)                     8.833760e+12     6.0  2.800113  0.010109
C(HOSP_REGION_DES):C(RACE_DES)  1.549081e+13    18.0  1.636752  0.043352
Residual                        3.481834e+15  6622.0       NaN       NaN
```

Kruskal Wallis

In [7]:

```
group_sizes = clean_data_anova.groupby(['HOSP_REGION_DES', 'RACE_DES']).size()
print(group_sizes)
```

```
HOSP_REGION_DES  RACE_DES
Midwest          Asian or Pacific Islander      40
                 Black                         305
                 Hispanic                       65
                 Missing                       312
                 Native American                 5
                 Other                          62
                 White                         638
Northeast        Asian or Pacific Islander      34
                 Black                         191
                 Hispanic                      129
                 Missing                       121
                 Native American                 2
                 Other                         157
                 White                         229
South            Asian or Pacific Islander      66
                 Black                        1042
                 Hispanic                      515
                 Missing                       253
                 Native American                13
                 Other                         275
                 White                        1117
West             Asian or Pacific Islander      74
                 Black                          76
                 Hispanic                      368
                 Missing                       127
                 Native American                 9
                 Other                          96
                 White                         329
dtype: int64
```

In [8]:

```
# Filter groups with at least 3 observations
filtered_data = clean_data_anova.groupby(['HOSP_REGION_DES', 'RACE_DES']).filter(lambda x: len(x) >= 3)
```

In [9]:

```
# Group by each factor and check if there are at least two groups
groups_by_region = filtered_data.groupby('HOSP_REGION_DES')
groups_by_race = filtered_data.groupby('RACE_DES')

# Kruskal-Wallis Test for 'HOSP_REGION_DES' if there are at least two groups
if len(groups_by_region) >= 2:
    kruskal_results_region = stats.kruskal(*[group['TOTCHG'] for name, group in groups_by_region])
    print('Kruskal-Wallis Test for HOSP_REGION_DES:', kruskal_results_region)
else:
    print('Not enough groups for Kruskal-Wallis Test on HOSP_REGION_DES')

# Kruskal-Wallis Test for 'RACE_DES' if there are at least two groups
if len(groups_by_race) >= 2:
    kruskal_results_race = stats.kruskal(*[group['TOTCHG'] for name, group in groups_by_race])
    print('Kruskal-Wallis Test for RACE_DES:', kruskal_results_race)
else:
    print('Not enough groups for Kruskal-Wallis Test on RACE_DES')
```

```
Kruskal-Wallis Test for HOSP_REGION_DES: KruskalResult(statistic=35.538455797703996, pvalue=9.3746200866394e-08)
Kruskal-Wallis Test for RACE_DES: KruskalResult(statistic=64.05097955034579, pvalue=6.738652359806292e-12)
```

In [10]:

```
# Dunn's Test for 'HOSP_REGION_DES'
dunn_test_region = sp.posthoc_dunn(filtered_data, val_col='TOTCHG', group_col='HOSP_REGION_DES', p_adjust='bonferroni')
print('Dunn\'s Test for HOSP_REGION_DES:\n', dunn_test_region)

# Dunn's Test for 'RACE_DES'
dunn_test_race = sp.posthoc_dunn(filtered_data, val_col='TOTCHG', group_col='RACE_DES', p_adjust='bonferroni')
print('Dunn\'s Test for RACE_DES:\n', dunn_test_race)
```

```
Dunn's Test for HOSP_REGION_DES:
                 Midwest  Northeast         South          West
Midwest    1.000000e+00   0.072759  1.000000e+00  8.501805e-07
Northeast  7.275934e-02   1.000000  2.144338e-01  1.365091e-01
South      1.000000e+00   0.214434  1.000000e+00  8.787442e-07
West       8.501805e-07   0.136509  8.787442e-07  1.000000e+00
Dunn's Test for RACE_DES:
                            Asian or Pacific Islander     Black      Hispanic  \
Asian or Pacific Islander                   1.000000  1.000000  1.000000e+00
Black                                       1.000000  1.000000  3.054186e-03
Hispanic                                    1.000000  0.003054  1.000000e+00
Missing                                     0.245207  0.180464  3.377957e-07
Native American                             0.284367  0.610279  6.873750e-02
Other                                       1.000000  0.000462  1.000000e+00
White                                       1.000000  0.007327  1.000000e+00

                                Missing  Native American         Other  \
Asian or Pacific Islander  2.452072e-01         0.284367  1.000000e+00
Black                      1.804643e-01         0.610279  4.617213e-04
Hispanic                   3.377957e-07         0.068737  1.000000e+00
Missing                    1.000000e+00         1.000000  9.452361e-08
Native American            1.000000e+00         1.000000  2.992629e-02
Other                      9.452361e-08         0.029926  1.000000e+00
White                      4.092657e-07         0.111623  1.000000e+00

                                  White
Asian or Pacific Islander  1.000000e+00
Black                      7.327373e-03
Hispanic                   1.000000e+00
Missing                    4.092657e-07
Native American            1.116230e-01
Other                      1.000000e+00
White                      1.000000e+00
```

<font : color = 'red'> Interpretation for 'HOSP_REGION_DES': </font> Midwest vs. Other Regions:

Midwest vs. Northeast: No significant difference (p ≈ 0.0728). Midwest vs. South: No significant difference (p = 1.0). Midwest vs. West: Significant difference (p ≈ 8.50 × 10^{-7}).

Northeast vs. Other Regions:

Northeast vs. South: No significant difference (p ≈ 0.2144). Northeast vs. West: No significant difference (p ≈ 0.1365). South vs. West: Significant difference (p ≈ 8.79 × 10^{-7}).

<font : color = 'blue'>Interpretation for 'RACE_DES':</font>

Asian or Pacific Islander vs. Other Races: No significant differences observed against any race.

Black vs. Other Races:

Black vs. Hispanic: Significant difference (p ≈ 0.0031). Black vs. Other: Significant difference (p ≈ 0.0005). Hispanic vs. Other Races:

Hispanic vs. Missing: Significant difference (p ≈ 3.38 × 10^{-7}). Missing vs. Other Races:

Missing vs. Other: Significant difference (p ≈ 9.45 × 10^{-8}). Missing vs. White: Significant difference (p ≈ 4.09 × 10^{-7}). Native American vs. Other: Significant difference observed against 'Other' (p ≈ 0.0299).

White vs. Other Races: No significant differences observed against any race except for a significant difference with 'Black' (p ≈ 0.0073).

Summary: In terms of hospital regions, significant differences in 'TOTCHG' are observed between the Midwest and West and between the South and West. Regarding race, several significant differences are observed, notably between Black and Hispanic, Black and Other, and between the Missing category and Hispanic and White. Where p-values are high (close to 1), it suggests no significant difference between those groups.

In [ ]: