EDA and Visualization Final

In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%config InlineBackend.figure_format = 'retina'
```

In [2]:

```
def set_spines(ax):
    ax.spines['top'].set_visible(False)
    ax.spines['right'].set_visible(False)
    ax.spines['left'].set_visible(True)
    ax.spines['bottom'].set_visible(True)
```

Data:

- Internet users by World Region: https://ourworldindata.org/internet (https://ourworldindata.org/internet)
- Internet Trends: https://www.kleinerperkins.com/perspectives/internet-trends-report-2018/ (https://www.kleinerperkins.com/perspectives/internet-trends-report-2018/)

Data Processing

In [3]:

```
years = [i for i in range(2008,2018,1)]
hours = [0,0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5]
laptop = [2.2, 2.3, 2.4, 2.6, 2.5, 2.3, 2.2, 2.2, 2.2, 2.1]
mobile = [.3,.3,.4,.8,1.6,2.3,2.6,2.8, 3.1,3.1]
other = [0.2,.3,.4,.3,.3,.3,.3,.4,.4,.6]
```

In [4]:

```
changes = {'lap_change':(2.2, 2.1), 'mob_change':(.3, 3.1),'oth_change':(.2, 0.6)}
```

In [5]:

```
reg = pd.read_csv('internet-users-by-world-region.csv')
```

In [6]:

```
reg = reg.rename(columns={'Internet Users by World Region (World Bank (2016))': "users"})
reg = reg[reg['Year']>=2000]
```

In [7]:

```
data = []
data = [reg[reg['Entity']==i] for i in reg.Entity.value_counts().index]
```

In [8]:

```
data[6].head()
```

Out[8]:

	Entity	Year	users
79	Middle East & North Africa	2000	5335063.5
80	Middle East & North Africa	2001	6669439.0
81	Middle East & North Africa	2002	12293885.0
82	Middle East & North Africa	2003	17099494.0
83	Middle East & North Africa	2004	28379684.0

In [10]:

```
data = [reg[reg['Entity']==i] for i in reg.Entity.value_counts().index]
```

In [11]:

```
df = pd.DataFrame(data={'laptop':[laptop[0], laptop[9]], 'mobile':[mobile[0],mobile[9]], 'other':[other[0],ot
her[9]]})
df = df.T
```

In [12]:

```
df['type'] = ['laptop', 'mobile','other']

changes
```

Out[12]:

```
{'lap_change': (2.2, 2.1), 'mob_change': (0.3, 3.1), 'oth_change': (0.2, 0.6)}
```

In [13]:

```
for i in data:
    print(f"{i['Entity'].iloc[0]}:   {i['users'].iloc[0]/1000000} , {i['users'].iloc[-1]/1000000}")
```

```
South Asia:   6.568352 , 412.109408
North America:   137.339792 , 271.351008
Latin America & Caribbean:   20.529908 , 344.699296
Sub-Saharan Africa:   3.3461643 , 224.100224
East Asia & Pacific:   114.411096 , 1135.598208
Europe & Central Asia:   113.651216 , 651.396608
Middle East & North Africa:   5.3350635 , 185.348464
```

In [14]:

```
reg.groupby('Entity').first().sort_values('users')
```

Out[14]:

	Year	users
Entity		
Sub-Saharan Africa	2000	3346164.3
Middle East & North Africa	2000	5335063.5
South Asia	2000	6568352.0
Latin America & Caribbean	2000	20529908.0
Europe & Central Asia	2000	113651216.0
East Asia & Pacific	2000	114411096.0
North America	2000	137339792.0

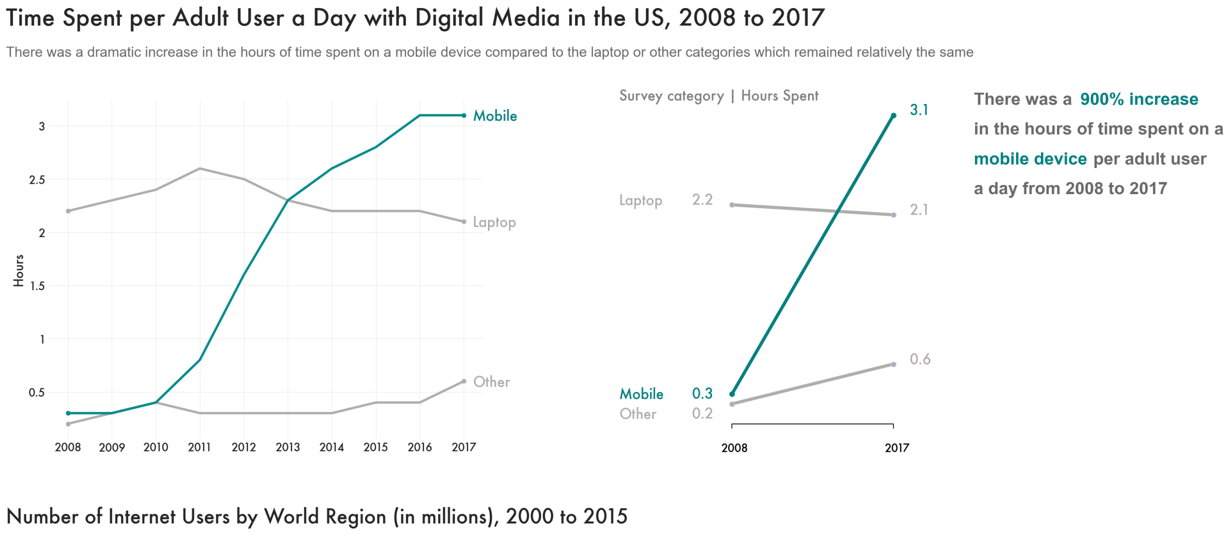
In [15]:

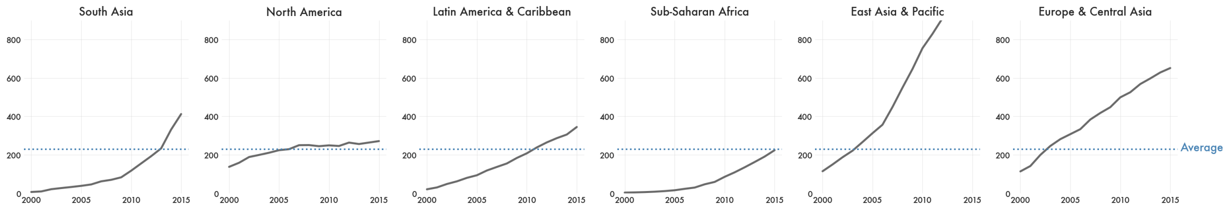
```
names = reg.groupby('Entity').first().sort_values('users').index.to_list()
```

In [16]:

```
# Styling
plt.style.use('seaborn-dark')

fig, ax = plt.subplots(1,2, figsize=(22,8))

ax[0].set_xticks([0,1,2,3,4,5,6,7,8,9,10,11,12,13])
ax[0].set_xticklabels(years, fontsize=14, fontname = 'Futura')
ax[0].set_yticks([0,0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5, 5.5])
ax[0].set_yticklabels(hours, fontsize=14, fontname = 'Futura')


ax[0].text(9.2,3.05,'Mobile',fontsize=18, fontweight='heavy', fontname='Futura', color = 'teal')
ax[0].text(9.2,2.05,'Laptop',fontsize=18, fontweight='heavy', fontname='Futura', color = '#b0adac')
ax[0].text(9.2,.55,'Other',fontsize=18, fontweight='heavy', fontname='Futura', color = '#b0adac')
ax[0].set_ylabel('Hours', fontsize=16,fontname='Futura')
laptop = [2.2, 2.3, 2.4, 2.6, 2.5, 2.3, 2.2, 2.2, 2.2, 2.1]
mobile = [.3,.3,.4,.8,1.6,2.3,2.6,2.8, 3.1,3.1]
other = [0.2,.3,.4,.3,.3,.3,.3,.4,.4,.6]
ax[0].plot(laptop, color = '#b0adac', linewidth=3, marker='o', ms=5, markevery=9)
ax[0].plot(other, color = '#b0adac', linewidth=3, marker='o', ms=5,markevery=9)
ax[0].plot(mobile, color = 'darkcyan', linewidth= 3, marker='o', ms=5, markevery=9)

ax[0].set_facecolor('white')
ax[0].grid(color='lightgrey', linestyle='-', linewidth=.3)
set_spines(ax[0])
set_spines(ax[1])


# Create line plots
for i in changes:
    ax[1].plot([.5,1], [changes[i][0],changes[i][1]], color='#b0adac', marker='o', markeredgecolor='lightsteelblue', linewidth=4)
ax[1].plot([.5,1], [.3,3.1], color='teal', marker='o', markeredgecolor='teal', linewidth=4)

# Create year line
ax[1].plot([.5,1], [0,0], color='black', linewidth=1, marker=3)

# Set x and y limits
ax[1].set_xlim([0,1.35])


# Add text
xy = [[1.25,3.25],[1.58,3.25],[1.25,2.95],[1.25,2.65],[1.62,2.65],[1.25,2.35]]
text = ['There was a','900% increase','in the hours of time spent on a','mobile device','per adult user','a day from 2008 to 2017']
color = ['dimgrey', 'teal', 'dimgrey','teal','dimgrey','dimgrey']
for i in range(6):
    ax[1].text(xy[i][0], xy[i][1], text[i], ha='left', va='center', fontweight='bold', fontsize=22, color = color[i])

ax[1].text(0.55, -.25, '2008', horizontalalignment='right', verticalalignment='center', fontname='Futura',  fontsize=14, color='black')
ax[1].text(1.05, -.25, '2017', horizontalalignment='right', verticalalignment='center', fontname='Futura',  fontsize=14, color='black')

xy = [[0.15,3.25],[.15,2.2],[.15,.25],[.15,.05],[.375,2.2],[.375,.25],[.375,0.05],[1.05,2.1],[1.05,3.1],[1.05,.6]]
text =['Survey category | Hours Spent','Laptop','Mobile','Other','2.2','0.3','0.2','2.1','3.1','0.6']
color = ['grey','#b0adac','teal','#b0adac','#b0adac','teal','#b0adac','#b0adac','teal','#b0adac']
fw=['normal','normal','heavy','normal','normal','heavy','normal','normal','heavy','normal']
for i in range(10):
    ax[1].text(xy[i][0], xy[i][1], text[i], ha='left', fontname='Futura', fontweight=fw[i], fontsize=18, color = color[i])

# Set background to white
ax[1].set_facecolor('xkcd:white')
# Remove x-axis
ax[1].set_xticks(range(0))
ax[1].set_yticks(range(0))
ax[1].set_yticklabels([])
```

```
plt.text(-1.75,4,'Time Spent per Adult User a Day with Digital Media in the US, 2008 to 2017', fontname = 'Fu
tura', fontsize=30)
plt.text(-1.75, 3.725, 'There was a dramatic increase in the hours of time spent on a mobile device compared
 to the laptop or other categories which remained relatively the same',
                         horizontalalignment='left', verticalalignment='center',  fontname='Arial', fontsize=
18, color='dimgrey')

plt.text(-1.75,-1,'Number of Internet Users by World Region (in millions), 2000 to 2015', fontsize=25, fontna
me = 'Futura')

#plt.show()
#plt.tight_layout()

fig, ax = plt.subplots(1,6, figsize=(26,4))
ax = ax.flatten()

for i in range(6):
    ax[i].plot(data[i]['Year'], data[i]['users']/1000000, color ='dimgrey', linewidth=2.5)
    ax[i].plot(data[i]['Year'], data[i]['users']/1000000, color ='dimgrey', linestyle = 'dashed', linewidth=
1.5)

    ax[i].set_facecolor('white')
    ax[i].grid(color='lightgrey', linestyle='-', linewidth=.3)
    ax[i].axhline(reg['users'].mean()/1000000, color='steelblue',linestyle='dotted', linewidth=2)


    ax[i].set_title(reg.Entity.value_counts().index[i], fontsize=14, fontname='Futura', loc='center')
    ax[i].set_xticks(range(2000,2016,5))
    ax[i].set_xticklabels(range(2000,2016,5), fontsize=10, fontname = 'Futura')
    ax[i].set_yticks(range(0,900,200))
    ax[i].set_yticklabels(range(0,900,200), fontsize=10, fontname = 'Futura')
    ax[i].set_ylim(0,900)

    set_spines(ax[i])
plt.text(2016,220,'Average', fontsize=14,color='steelblue', fontname='Futura',clip_on=False)

plt.show()
```





Data:

- Insurance Data: [https://www.kaggle.com/mirichoi0218/insurance (https://www.kaggle.com/mirichoi0218/insurance)](https://www.kaggle.com/mirichoi0218/insurance)

In [17]:

```
df = pd.read_csv('insurance.csv')
```

In [18]:

```
df.head()
```

Out[18]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

In [19]:

```
se = df[df['region']=='southeast']
sw = df[df['region']=='southwest']
ne = df[df['region']=='northeast']
nw = df[df['region']=='northwest']
```

In [20]:

```
sm = df[df['smoker']=='yes']
nsm = df[df['smoker']=='no']
```

In [23]:

```
noch.shape
```

Out[23]:

```
(574, 7)
```

In [24]:

```
sm = df[df['smoker']=='yes']
nsm = df[df['smoker']=='no']
```

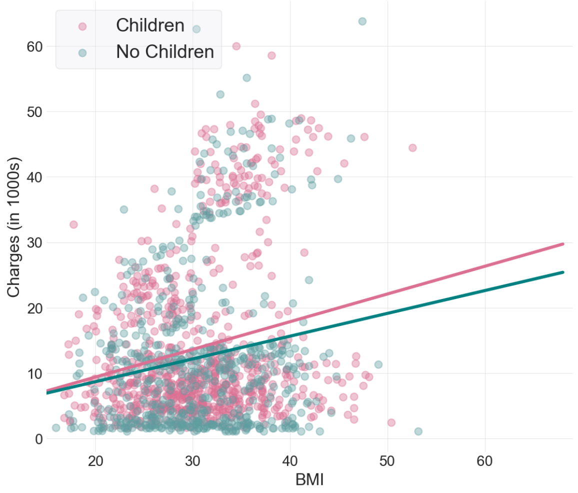
In [27]:

```
bmi = df['bmi']
charges = df['charges']

fig, ax = plt.subplots(1,2, figsize=(20,8))

plt.text(-30,80, 'Visualizing Health Insurance Costs', fontsize=40, fontweight='bold', fontname='Arial', colo
r = 'darkslategrey')
ax[1].text(10,70, 'BMI vs. Health Insurance Costs', fontsize=25, fontweight='bold', fontname='Arial', color =
'slategrey')
ax[1].set_xticklabels(range(15,55,5), fontsize=14)
ax[1].set_yticklabels(range(-10,65,10), fontsize=14)
ax[1].set_xlabel('BMI', fontsize=16)
ax[1].set_ylabel('Charges (in 1000s)', fontsize=16)

ax[1].scatter(sm['bmi'], sm['charges']/1000, s=50, c='xkcd:dusty pink', alpha = 0.4, label = 'Smoker')
m, b = np.polyfit(sm['bmi'], sm['charges']/1000, 1)
x=np.linspace(15,50)
ax[1].plot(x, m*x + b, color = 'xkcd:darkish pink',linewidth=3)


ax[1].scatter(nsm['bmi'], nsm['charges']/1000,  s = 50, c='xkcd:dusty blue', alpha = 0.4, label = 'Non Smoke
r')
m, b = np.polyfit(nsm['bmi'], nsm['charges']/1000, 1)
x=np.linspace(15,50)
ax[1].plot(x, m*x + b, color = 'xkcd:dusty blue', linewidth=3)

ax[1].legend(framealpha=.3, frameon=True, prop={'size': 18})
ax[1].set_facecolor('white')
ax[1].grid(color='lightgrey', linestyle='-', linewidth=.3)
ax[1].set_xlim([15,50])
set_spines(ax[1])


ax[0].text(10,70, 'BMI vs. Health Insurance Costs', fontsize=25, fontweight='bold', fontname='Arial', color =
'slategrey')
ax[0].set_xticklabels(range(10,70,10), fontsize=14)
ax[0].set_yticklabels(range(-10,65,10), fontsize=14)
ax[0].set_xlabel('BMI', fontsize=16)
ax[0].set_ylabel('Charges (in 1000s)', fontsize=16)

ax[0].scatter(ch['bmi'], ch['charges']/1000, s=50, c='palevioletred', alpha = 0.4, label = 'Children')
m, b = np.polyfit(ch['bmi'], ch['charges']/1000, 1)
x=np.linspace(15,68)
ax[0].plot(x, m*x + b, color = 'palevioletred',linewidth=3)


ax[0].scatter(noch['bmi'], noch['charges']/1000,  s = 50, c='cadetblue', alpha = 0.4, label = 'No Children')
m, b = np.polyfit(noch['bmi'], noch['charges']/1000, 1)
x=np.linspace(15,68)
ax[0].plot(x, m*x + b, color = 'teal', linewidth=3)

ax[0].legend(framealpha=.3, loc='upper left',frameon=True, prop={'size': 18})
ax[0].set_facecolor('white')
ax[0].grid(color='lightgrey', linestyle='-', linewidth=.3)
ax[0].set_xlim([15,69])
set_spines(ax[0])

#              )
#plt.colorbar(m)
bmi = df['bmi']
charges = df['charges']

fig, ax = plt.subplots(1,2, figsize=(20,8))

ax[1].text(10,70, 'Age vs. Health Insurance Costs', fontsize=25, fontweight='bold', fontname='Arial', color =
'slategrey')
ax[1].set_xticklabels(range(15,55,5), fontsize=14)
ax[1].set_yticklabels(range(-10,65,10), fontsize=14)
ax[1].set_xlabel('Age', fontsize=16)
ax[1].set_ylabel('Charges (in 1000s)', fontsize=16)

ax[1].scatter(sm['age'], sm['charges']/1000, s=50, c='xkcd:dusty pink', alpha = 0.4, label = 'Smoker')
```

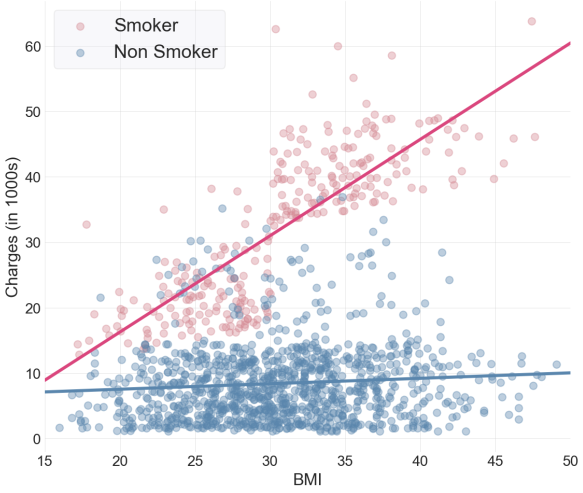
```
m, b = np.polyfit(sm['age'], sm['charges']/1000, 1)
x=np.linspace(15,50)
ax[1].plot(x, m*x + b, color = 'xkcd:darkish pink',linewidth=3)


ax[1].scatter(nsm['age'], nsm['charges']/1000,  s = 50, c='xkcd:dusty blue', alpha = 0.4, label = 'Non Smoker')
m, b = np.polyfit(nsm['age'], nsm['charges']/1000, 1)
x=np.linspace(15,50)
ax[1].plot(x, m*x + b, color = 'teal', linewidth=3)

ax[1].legend(framealpha=.3, frameon=True, prop={'size': 18})
ax[1].set_facecolor('white')
ax[1].grid(color='lightgrey', linestyle='-', linewidth=.3)
ax[1].set_xlim([15,50])
set_spines(ax[1])


ax[0].text(10,70, 'Age vs. Health Insurance Costs', fontsize=25, fontweight='bold', fontname='Arial', color = 'slategrey')
ax[0].set_xticklabels(range(10,70,10), fontsize=14)
ax[0].set_yticklabels(range(-10,65,10), fontsize=14)
ax[0].set_xlabel('Age', fontsize=16)
ax[0].set_ylabel('Charges (in 1000s)', fontsize=16)

ax[0].scatter(ch['age'], ch['charges']/1000, s=50, c='palevioletred', alpha = 0.4, label = 'Children')
m, b = np.polyfit(ch['age'], ch['charges']/1000, 1)
x=np.linspace(15,68)
ax[0].plot(x, m*x + b, color = 'palevioletred',linewidth=3)


ax[0].scatter(noch['age'], noch['charges']/1000,  s = 50, c='cadetblue', alpha = 0.4, label = 'No Children')
m, b = np.polyfit(noch['age'], noch['charges']/1000, 1)
x=np.linspace(15,68)
ax[0].plot(x, m*x + b, color = 'teal', linewidth=3)

ax[0].legend(framealpha=.3, loc='upper left',frameon=True, prop={'size': 18})
ax[0].set_facecolor('white')
ax[0].grid(color='lightgrey', linestyle='-', linewidth=.3)
ax[0].set_xlim([15,69])
set_spines(ax[0])

#               )
#plt.colorbar(m)
```

Visualizing Health Insurance Costs

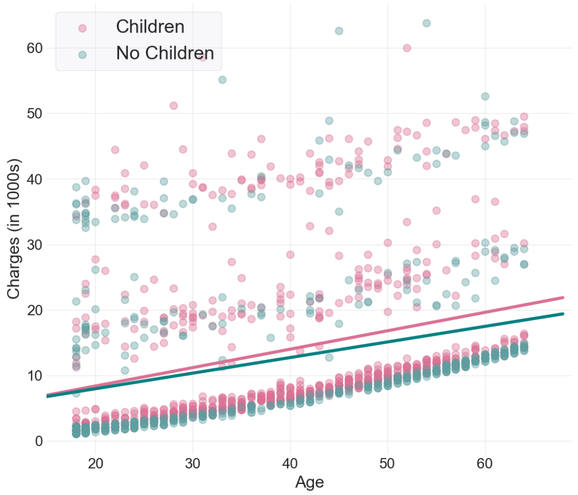
BMI vs. Health Insurance Costs



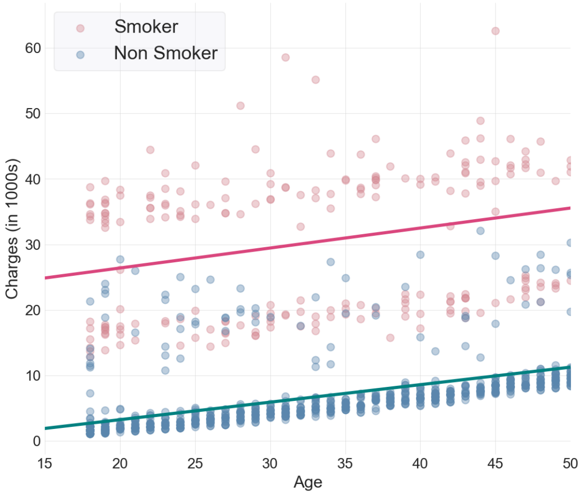
BMI vs. Health Insurance Costs



Age vs. Health Insurance Costs



Age vs. Health Insurance Costs


Smoker
Non Smoker
Charges (in 1000s)
Age